مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية
King Abdullah Bin Abdulaziz Int'l Center for The Arabic Language

المُعالَجة الآليَّة للنُّصُوص العربيَّة

مباحث لغوية ٥٧

تحرير

د. مُحسن رَشوان   د. المُعتزّ بالله السَّعيد

الباحثون:

د. وليد مجدي   د. أسامة إمام
د. أحمد رافع   د. مُحسن رَشوان
د. علي فهمي

مباحث لغوية ٥٧

المُعالَجة الآليَّة للنُّصُوص العربيَّة

تحرير:

د. مُحسن رَشوان | د. المُعتزّ بالله السَّعيد

الباحثون:

د. وليد مجدي | د. أسامة إمام

د.أحمد رافع | د. مُحسن رَشوان

د. علي علي فهمي

١٤٤١هـ - ٢٠١٩م

مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية
King Abdullah Bin Abdulaziz Int'l Center for The Arabic Language

مركز الملك عبدالله بن عبدالعزيز الدولي
لخدمة اللغة العربية
King Abdullah Bin Abdulaziz Int'l Center for
The Arabic Language

المُعالَجة الآليَّة للنُّصُوص العربيَّة

الطبعة الأولى
١٤٤١ هـ - ٢٠١٩ م

المملكة العربية السعودية - الرياض
ص.ب ١٢٥٠٠ الرياض ١١٤٧٣
هاتف:٠٠٩٦٦١١٢٥٨٧٢٦٨ - ٠٠٩٦٦١١٢٥٨١٠٨٢
البريد الإليكتروني: nashr@kaica.org.sa

مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة
العربية، ١٤٤١هـ.
فهرسة مكتبة الملك فهد الوطنية أثناء النشر
رشوان، محسن
المعالجة الآلية للنصوص العربية./ محسن رشوان؛ المعتز بالله
السعيد .-الرياض، ١٤٤٠هـ
..ص؛ .. سم
ردمك: ٥ -٥٢ -٨٢٢١ -٦٠٣ -٩٧٨
١ - اللغة العربية - معالجة البيانات أ. السعيد ، المعتز بالله
(مؤلف مشارك) ب. العنوان
ديوي ٤١٠,٢٨٥ ١٤٤٠/١٠١٦٦
رقم الإيداع: ١٤٤٠/١٠١٦٦
ردمك: ٥ -٥٢ -٨٢٢١ -٦٠٣ -٩٧٨

التصميم والإخراج
دار وجوه للنشر والتوزيع
Wojooh Publishing & Distribution House
www.wojoooh.com
المملكة العربية السعودية - الرياض
الهاتف:4562410 الفاكس:4561675
للتواصل والنشر:
info@wojoooh.com

بسم الله الرحمن الرحيم


مَكْتَبَةُ
لِسَانُ العَرَب
أ. علاء الدين شوقي
@lisanarb
www.lisanarb.com

فهرس الكتاب

كلمة المركز

يعمل المركز في مجال البحث العلمي ونشر الكتب مستهدفاً التركيز على المجالات البحثية التي ما زالت بحاجة إلى تسليط الضوء عليها، وتكثيف البحث فيها، ولفت أنظار الباحثين والجهات الأكاديمية إلى أهمية استثمارها بمختلف وجوه الاستثمار، وذلك مثل مجال (التخطيط اللغوي) و (العربية في العالم) و(الأدلة والمعلومات) و (تعليم العربية لأبنائها أو لغير الناطقين بها) إلى غير ذلك من المجالات، وإن من أهم مجالات البحث المستقبلية في اللغة العربية مجال (العربية والحوسبة ، والذكاء الاصطناعي) حيث إن حياة اللغات ومستقبلها مرهونة بمدى تجاوبها مع التطورات التقنية والعالم الافتراضي، وكثافة المحتوى الالكتروني المكتوب، وهو ما يشكّل تحديا حقيقيا أمام اللغات غير المنتجة للمعرفة أو للتقنية.

وقد عمل المركز على تسليط الضوء على هذا المجال التخصصي؛ مستعينا بالكفاءات القادرة من المهتمين بالتخصص البيني (بين اللغة والحاسوب) مقدّرا جهودهم، وهادفاً إلى نشرها، وتعميم مبادئها، راغباً أن يكون هذا المسار العلمي مقررا في الجامعات في كلية العربية والحاسوب، ومجالا بحثيا يقصده الباحثون الأكديميون، والجهات البحثية العربية.

وقد أصدر المركز سابقا ستة عشر كتاباً مختصا في (حوسبة العربية) وفي الإفادة من (المدونات اللغوية) في الأبحاث العربية، ويحتفل بإصدار سبعة كتب جديدة مختصة في (حوسبة العربية والذكاء الاصطناعي)، ويقدمها للقارئ العربي، وللجهات الأكاديمية؛ للإفادة منها في مناهج التعليم والبناء عليه، وهذه الكتب السبعة هي: (العربيَّة والذَّكاء الاصطناعي، تطبيقات الذكاء الاصطناعي في خدمة اللغة العربية، خوارزميات الذكاء الاصطناعي في تحليل النص العربي، مقدمة في حوسبة اللغة العربية، الموارد اللغوية الحاسوبية، المعالجة الآلية للنصوص العربية، تطبيقات أساسية في المعالجة الآلية للغة العربية).

ويشكر المركز السادة مؤلفي الكتب، ومحرريها، لما تفضلوا به من عمل علمي رصين، وأدعو الباحثين والمؤلفين إلى التواصل مع المركز لاستكمال المسيرة، وتفتيق فضاءات المعرفة.

وفق الله الجهود وسدد الرؤى.

الأمين العام

أ. د. محمود إسماعيل صالح

مقدمة

تخضعُ اللُّغاتُ الطَّبيعيَّةُ لعدَّة مُستوياتٍ في الـمُعالجَة الآليَّة. وتتدرَّجُ من مُستوياتٍ تُعالِجُ البنيةَ السَّطحيَّةَ إلى مُستوياتٍ أخرى تُعالجُ البنيةَ العميقة. والواقعُ أنَّ نطاقَ طُمُوح الباحثينَ في ميادين حوسَبة اللُّغة يتَّسعُ مع الطَّفرات الهائلة الَّتي يشهدُها عالَمُ الذَّكاء الاصطناعيّ. ولم تَعُد الرُّؤى الاستشرافية للـمُستقبل قاصرةً على تمكين الآلة من فهم المجموعات المحدودة من النُّصُوص؛ بل تجاوزَت ذلكَ إلى رغبةٍ حقيقيَّةٍ في تمكين الآلة من التَّعامُل مع مجموعاتٍ كبيرةٍ نسبيًّا من النُّصُوص الـمُتَضَمَّنة في الذَّخائر اللُّغويَّة ومُستودَعات البيانات.

وتُعَدُّ اللُّغةُ العربيَّةُ إحدى اللُّغات الطَّبيعيَّة الَّتي تنالُ حظًّا وافرًا من عناية الباحثينَ في حَوسَبة اللُّغة وتقنياتها؛ سواءٌ في صُورتِها المنطوقة أم المكتوبة؛ وسواءٌ على مُستوى محارفها ومبانيها، أم على مُستوى تراكيبها ومعانيها. ومع التَّطوُّر الملموس في الـمُعالجَة الآليَّة للُّغة العربيَّة، فإنَّنا نعتقدُ أنَّ الـمُستقبلَ القريبَ قد يشهدُ إجابةً عن بعض التَّساؤلات الَّتي لا تزالُ مطروحةً بشأن قواعد العربيَّة وماهيَّتِها وقوانين تطوُّرها وأنماطِها التَّركيبيَّة والدِّلاليَّة.

إنَّنا نُقَدِّمُ اليومَ هذا الكتابَ (الـمُعالَجَة الآليَّة للنُّصُوص العربيَّة) الَّذي يُعَدُّ الكتابَ الثَّالثَ ضمنَ سلسلة دراسات وبُحُوث في حَوسَبَة اللُّغة العربيَّة. وترتكزُ مادَّةُ هذا الكتاب على نُصُوص اللُّغة العربيَّة، وطرائق مُعالجتِها آليًّا. ونسعى من خلال ذلكَ إلى تقديم تصوُّرٍ حولَ واقع مُعالجة النَّصّ العربيّ من ناحية، وتوجيه القارئ الكريم إلى أبرز التَّحدِّيات في ذلكَ الـمَيدان من ناحيةٍ ثانية، والدَّفع برُؤيةٍ طموحةٍ للمأمول بشأن مُعالَجَة النُّصُوص العربيَّة من ناحيةٍ ثالثة.

ورغبةً في تحقيق أهدافِنا المنشودة، فقد قسَّمنا الكتابَ إلى أربعة فُصُولٍ، على النَّحو الآتي:

- الفصل الأوَّل: استرجاع المعلومات؛ يُعنى بمفهوم استرجاع المعلومات، وآليَّات العثور على مُحتوًى مُعَيَّنٍ في مجموعةٍ كبيرةٍ من الوثائق، لا سيَّما الوثائق النَّصِّيَّة. ويعرضُ للتَّمييز بينَ خاصِّيَّتَي البحث Search والتَّحرِّي Find؛ ويُعنى الفصلُ بمُحرِّكات البحث وهياكلها ووظائفِها وأساليب تطويرها.
- الفصل الثاني: التَّرجمة الآليَّة؛ ويَعرِضُ مُقدِّمةً مُوجَزةً حولَ التَّرجمة الآليَّة وأهمّ الـمُصطلحات الـمُستَخدَمة في ذلك الميدان؛ ويعرضُ كذلكَ لتقنيات التَّرجمة الآليَّة، والتَّوجُّهات البحثيَّة لتطويرها، والأدوات والموارد الأساسيَّة فيها. ويُقدِّمُ الفصلُ مجموعةً من الأفكار البحثيَّة الـمُوجِّهة لبناء موارد التَّرجمة الآليَّة.
- الفصل الثَّالث: التَّشكيل الآليّ؛ يُعنى هذا الفصلُ بآليَّة تشكيل النُّصُوص العربيَّة؛ ويُقدِّمُ تعريفًا بعلامات الضَّبط العربيَّة، كما يُقدِّمُ صياغةً رياضيَّة قياسيَّة لـمُعالَجَة إشكالات التَّشكيل. ويعرضُ الفصلُ أيضًا لأبرز الأساليب الـمُستخدَمة في تطوير آليَّة تشكيل النُّصُوص العربيَّة، والموارد اللَّازمة لذلك؛ ويعرضُ أخيرًا لبعض الأفكار البحثيَّة الَّتي يُمكنُ استثمارُها في إعداد أطروحاتٍ علميَّةٍ مُستقبليَّة.

- الفصل الرَّابع: التَّنقيب في النُّصُوص؛ ويشتملُ على ثلاثة مباحث؛ حيثُ يُقدِّمُ في المبحث الأوَّل لأساليب تجميع النُّصُوص وتصنيفها، والتَّطبيقات العمليَّة للتَّجميع والتَّصنيف في العربيَّة؛ ويُعنى المبحثُ الثَّاني بتلخيص النُّصُوص وأنواعِهِ وأساليبِه ونماذج أنظمتِه. أمَّا المبحثُ الثَّالثُ فيعرضُ لتطبيق استنباط اتِّجاهات الرَّأي العامّ، الَّذي يُعَدُّ أحدَ أبرز تطبيقات التَّنقيب في النُّصُوص. ويعرضُ هذا المبحثُ الأخيرُ لأساليب التَّنقيب عن الآراء وطرائق ذلكَ في اللُّغة العربيَّة، والموارد اللُّغويَّة اللَّازمة؛ كما يُقدِّمُ رؤيةً للتَّوجُّهات الـمُستقبليَّة والتَّحدِّيات الَّتي تُواجهُ التَّنقيب عن الآراء.

وبعدُ؛ فالكتابُ خُطوةٌ على الطَّريق إلى حوسَبة النُّصُوص العربيَّة وتيسير مُعالجتِها آليًّا. ونحنُ ننشُدُ أن تلي هذه الخُطوةَ خُطواتٌ أخرى أكثرُ عُمقًا وإدراكًا لبنية النُّصُوص العربيَّة، سعيًا إلى مُعالجة إشكالات هذه النُّصُوص، وابتكار أساليبَ جديدة وناجعة لتحسين نتائج مُخرَجاتِها.

نسألُ الله تعالى أن يتقبَّلَ هذا الجهدَ بالذِّكر الحَسَنِ والأجرِ الجزيل، وأن يجعلَه من العلمِ الَّذي ينفعُ أصحابَه بعد مماتهم.

ربَّنا عليكَ توكَّلنا وإليكَ أنبنا وإليكَ المصِير.

الـمُحرِّران

الفصل الأول
استِرجاع المعلُومات

د. وَليد مَجدي   د. أسامة إمام

١ – مقدمة.

٢ – عملية الفهرسة.

٣– آليَّة البحث.

٤ – تقييم البحث.

٥–محركات بحث الشَّبكة العنكبوتيَّة.

٦– محركات البحث المكتبية.

٧– محركات بحث الشبكات الاجتماعية.

٨– البحث الدلالي.

٩– أفكار تصلح للأطروحات العلميَّة (الماجستير والدكتوراه).

١٠– مِن المواقع الإلكترونيَّة التَّعليميَّة والإرشاديَّة.

١ – مقدمة

١,١ – تعريف استرجاع المعلومات

يُعَرَّفُ استرجاعُ المعلومات (Retrieval Information) بأَنَّهُ إيجاد محتوى غالبًا ما يكون وثائق من وسط مجموعة كبيرة من الوثائق ذات الطبيعة الغير منظمة، بحيث تقوم هذه الوثائق المسترجعة بسدّ الاحتياجات المعلوماتية للمستخدم. وبعبارةٍ أخرى، استرجاع المعلومات هو فن البحث عن المعلومات ذات الصلة بالموضوع الذي يبحث عنه المستخدم. وقد أصبحَ استرجاع المعلومات أحد أهم عناصر التطور المعلوماتيّ نتيجةً للزيادة المطردة في كمية المعلومات المتوافرة، والتي تتزايد باستمرار مما يجعل أهمية الوصول لهذه المعلومات بطريقة ممنهجة وسريعة من أهم متطلبات أي منظومة معلوماتية قوية.

١,٢ – الفرق بين خاصِّيَّتَي البحث (Search) والتَّحَرِّي (Find)

هناك فرق كبير بين استرجاع المعلومات الذي يعتمد على البحث (search) وذلك الَّذي يعتمد على خاصية التَّحرِّي (find) التي تتواجد في معظم برامج وتطبيقات الحاسوب، والتي تساعد المستخدم على تحديد كلمة في النص أو الصفحة التي يقرأها. فعند البحث عن كلمة ما في إحدى الوثائق أو في مجموعة من الوثائق عن طريق التَّحرِّي فإن مُعالِجَ الحاسوب يقوم بمطابقة كل كلمة في النص بطريقة متسلسلة ويتم تحديد مكان أو أماكن التطابق للمستخدم. هذه الطريقة في البحث يعيبها البطء في مُعالجَة عملية البحث، حيث يتم البحث بمطابقة كلمة كلمة في النص لكلمة البحث مما يجعل البحث في كمية كبيرة من النصوص والمستندات غير عملي لطول الوقت المطلوب لمطابقة كل الكلمات. ومع هذا تظل خاصية «التَّحَرِّي» من أهم الخصائص لمعظم تطبيقات الحاسوب لأنها تساعد المستخدم في تحديد بعض الكلمات في النص المقروء بطريقة سهلة وسريعة بالنسبة للنصوص الصغيرة نسبيًّا.

وفيما يتعلَّقُ باسترجاع المعلومات عن طريق البحث، فإن الموضوع يعتمد على طرق وعناصر مختلفة من أجل تحديد الوثائق المراد البحث عنها بطريقة أكثر عملية وبدقة وإمكانيات أعلى في علمية التطابق والبحث كما سيتضح فيما يلي.

١ , ٣- نظم استرجاع المعلومات (محركات البحث)

كما ذكرنا آنفًا، فإن استرجاع المعلومات عملية متكاملة وأكثر تعقيدا من مجرد استخدام التطابق المتسلسل لكلمة البحث مع النصوص كما هو الحال في خاصية «التَّحرِّي». هناك نظم كاملة لاسترجاع المعلومات تكون مسئولة بشكل أساسي عن استرجاع كل ما كانَ ذا صلة بما يبحث عنه المستخدم بطريقة دقيقة وسريعة. الاسم الشائع لنظم استرجاع المعلومات هو «محركات البحث». ويتكوَّن محرك البحث من مجموعة عناصر أساسية تقوم على معالجة الوثائق وموضوعات البحث بـطرق مختلفة من أجل الحصول على نتائج بحث مرضية للمستخدم. وتختلف كيفية معالجة البيانات والوثائق من تطبيق إلى آخر ومن لغة إلى أخرى؛ فمحركات بحث المكتبات تختلف من حيث معالجة المعلومات وطريقة البحث عن محركات بحث الإنترنت أو الويب؛ كما أن محرك بحث التطبيق الواحد يختلف من حيث طريقة المعالجة على حسب اللغة أو نوع البيانات التي يتم البحث بها. كمثال لهذا: مطابقة كلمة «احمد» و«أحمد» تحتاج إلى طريقة معالجة خاصة باللغة العربية، فيما أن معالجات مختلفة تكون مطلوبة للغات الأخرى ذات الخصائص المختلفة. طرق المعالجة وطريقة البحث وأسلوب عرض النتائج أهم وظائف محركات البحث، وهي التي تجعلها مختلفة تماما عن خاصية «التَّحرِّي» البسيطة التي تستخدم لتحديد بعض الكلمات أثناء القراءة.

١ , ٤- مجموعات المستندات والوثائق (ما يتم البحث بداخله)

المهمة الأساسية لمحرك البحث هي استرجاع الوثائق والمستندات ذات الصلة بما يبحث عنه المستخدم من أجل إشباع حاجته المعلوماتية. قد يُتَصَوَّر من الوهلة الأولى أن هذه الوثائق تكون وثائق نصية فقط، ولكن – في الحقيقة – استرجاع المعلومات يشمل أيَّ نوع من المعلومات بحيث تأخذ الوثائق صورًا مختلفة، فيمكن أن تكون ملفات نصية بسيطة، أو ملفات نصيَّة متقدمة كصفحات الويب، أو ملفات نصية منظمة كالملفَّات النَّصِّيَّة Words وملفات XML. وأيضًا يمكن أن تكون الوثائق غير نصية بالأساس، كالصور والملفات الصوتية والمرئيَّات. يمكن أن تكون مجموعة الملفات التي يتم البحث فيها كلها من نفس النوع أو من أنواع مختلفة مثلما يحدث في محركات بحث الويب، حيث تشتمل النتائج على صفحات ويب بالإضافة إلى صور ومرئيَّات.

تحتوي مجموعة الوثائق – أيًّا كان نوعها – في الغالب على أعداد كبيرة وهائلة من الوثائق حيث تصل إلى آلاف وملايين، بل ومليارات الوثائق كما هو الحال في الشَّبكة العنكبوتيَّة. ولهذا، فمن الضروري عند تطوير محركات البحث أن تكون قادرة على معالجة هذه الأعداد الهائلة بدقة وفي وقت سريع جدًّا. وكمثال على هذا، عند استخدام أحد محركات بحث الويب (مثل: Google)، فإن عملية البحث تتم في بضعة أجزاء من الثانية وبدقة عالية.

في بعض الأحيان يتوجب على محرك البحث أن يحدد تعريف الوثيقة التي يجب استرجاعها. فالوثيقة أحيانا لا تكون واضحة التعريف ومن هنا يكون تحديد عنصر الوثيقة (وحدة البحث) من واجبات محرك البحث.

أحد الأمثلة على هذا «محركات بحث المكتبة»؛ فأحد الخيارات أن تكون وحدة البحث هي الكتاب حيث تكون الوثائق المسترجعة في نتائج البحث هي قائمة بأسماء الكتب ذات الصلة.

كما يمكن أن تعرف الوثائق بأنها الفصول في الكتب أو الصفحات أو حتى الفقرات داخل الصفحة، بحيث تكون نتائج البحث عبارة عن قائمة بعناوين الفصول داخل بعض الكتب أو أرقام الصفحات أو الفقرات التي تحتوي على المعلومة المطلوبة.

في كل هذه الحالات يوجد نفس المستندات والمحتوى، ولكن تختلف طريقة تعريف عنصر الوثيقة وكيفية البحث وعرض النتائج.

١,٥- احتياجات المستخدم (المطلوب البحث عنه)

تختلفُ احتياجات المستخدم في عملية البحث من تطبيق بحث لآخر ومن شخصية لأخرى. فالسيناريو المعهود في عمليات استرجاع المعلومات والبحث أن يفكر المستخدم في موضوع ما ويحتاج إلى بعض المعلومات عنه، فيقوم بالتعبير عن هذا الموضوع ببضع كلمات ثم يقوم بالبحث عمَّا يريد. أحيانا تكون نتائج البحث غير مرضية بالنسبة للمستخدم، فيقوم بتغيير بعض كلمات البحث أو حتى إعادة صياغة الموضوع المراد البحث عنه بكلمات مختلفة كلية. هذا يوضح الفارق الأساسي بين شيئين في عملية استرجاع المعلومات، ألا وهما: موضوع البحث وكلمات البحث. يمكن تعريف موضوع

البحث بأنه ما يدور في خلد المستخدم عما يريد أن يجده؛ أما كلمات البحث فهي الكلمات المستخدمة للتعبير عن هذا الموضوع، وهي ليست بالضرورة أحسن ما يعبر عن هذا الموضوع. ويُبَيِّنُ المثال الموضح أسفله بعض الصياغات المختلفة لنفس موضوع البحث، إذ لا توجد بينها كلمة مشتركة واحدة. وإن طلبنا من أشخاص مختلفين صياغة كلمات بحث لنفس الموضوع، فمن الصعب أن نجد اثنين يصيغان نفس كلمة البحث. هذا يوضح إحدى الخصائص المهمة الواجب توافرها في أي محرك بحث فعال، حيث يفضل أن يتم البحث على مطابقة الموضوع، لا على المطابقة الحرفية للكلمات.

موضوع البحث: يريد المستخدم أن يعرف بعض المعلومات عن الهجمات على بُرجي التجارة العالميين في الولايات المتحدة الأمريكية عام ٢٠٠١.

بعض الصياغات الممكنة لموضوع البحث يمكن أن تكون كالآتي:

- أحداث ١١/ ٩.
- الهجمات على بُرجَي التجارة العالميين.
- الحوادث الإرهابية على الولايات المتحدة الأمريكية عام ٢٠٠١.
- تفجيرات ١١ سبتمبر – أمريكا.

وعلى النقيض لما تم توضيحه في المثال السابق، فإن موضوعات مختلفة يمكن أن تصاغ بنفس الكلمات مما يصنع بعض التخبط لمحرك البحث حيث لا يكون المقصود وراء كلمات البحث واضحًا تماما. ومثال هذا: قيام المستَخدِم بالبحث عن «محمد عبده». هنا موضوع البحث يحتمل احتمالات عدة للمقصود وراء كلمتي البحث كالآتي:

- محمد عبده: عالم دين مصري، عاش في أوائل القرن العشرين.
- محمد عبده: المطرب السعودي .
- محمد عبده يماني: وزير الثقافة السعوديّ في الفترة (١٣٩٥ هـ : ١٤٠٣هـ).
- محمد عبده صالح الوحش: اللاعب السَّابق في المنتخب المصري لكرة القدم.

وأمثلة أخرى كثيرة لهذا، مثل:

- «الرئيس الأمريكي جورج بوش»: (الأب أم الابن).
- «النادي الأهلي»: (المصري، السعودي، الليبي، القطري، الأردني، الإماراتي، أم البحريني).
- «الملك عبدالله»: (ملك السعودية، مؤسس الأردن، أم ملك الأردن الثاني).

كل هذه الأمثلة توضح أنه ليس بالضرورة أن تكون كلمات البحث معبرة بوضوح عن موضوع البحث، كما لا يلزم أن تكون لمستخدم محرِّكات البحث نتائج معينة متوقعة أو مُرضية لكل الأشخاص.

مما سبق يمكن استنتاج أن تعريف الاحتياجات المعلوماتية للمستخدم يختلف من شخص لآخر، وإن تشابه موضوع البحث أو حتى تشابهت كلمات البحث. وبالتالي فإن تعريف الوثائق المسترجعة التي تكون «ذات صلة» بموضوع البحث هو شيء نسبي غير محدد بالضرورة. ويُعدُّ هذا من أكبر التحديات التي تواجه أي محرك بحث من أهدافه أن يرضي المستخدمين عامة بتنوع توجهاتهم وأهدافهم.

١,٦- عملية استرجاع المعلومات

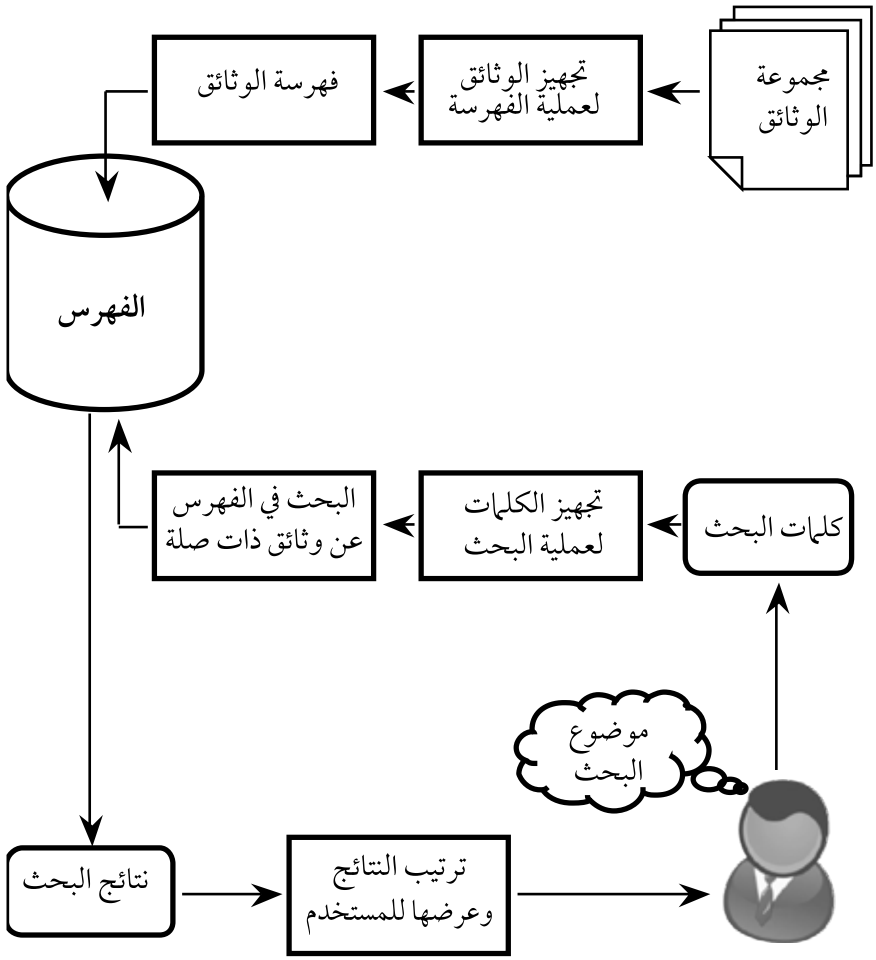
تتمُّ عملية استرجاع المعلومات على مرحلتين أساسيتين:

- المرحلة الأولى: مرحلة الفهرسة، حيث يتم تجهيز مجموعة الوثائق والمستندات المراد البحث فيها بالشكل المناسب وبناء الفهرس الرقمي للكلمات والمصطلحات من أجل تسهيل عملية البحث. هذه المرحلة من استرجاع المعلومات في بعض الأحيان تتم مرة واحدة فقط وبعدها تكون مجموعة الوثائق جاهزة للبحث، ولكن غالبا ما يُضطر إلى تكرار عملية الفهرسة كل فترة من الزمن، وهذا عند إضافة وثائق جديدة للمجموعة.
- المرحلة الأخرى: مرحلة البحث نفسها؛ وهي التي تتم بشكل متكرر كلما أراد أحد المستخدمين العثور على معلومات معينة. وتشمل هذه المرحلة أحيانا تجهيز كلمات البحث بشكل ما ليناسب البحث في الفهرس، ثم يتم البحث في الفهرس واسترجاع نتائج متماشية مع كلمات البحث ثم عرضها على المستخدم على شكل قائمة مرتبة حسب علاقة الوثيقة بموضوع البحث.

ويوضح الشكل (١-١) عملية استرجاع المعلومات بمرحلتيها.

كما هو مبين، هناك عمليات عدة لإتمام عملية استرجاع المعلومات، ولكن المستخدم لا يرى مما يحدث في الخلفية من هذه العملية إلا ما يكتبه من كلمة بحث ويعود من نتائج في النهاية.

وبعبارةٍ أخرى، لا يتعرَّف المستخدم على نوع المعالَجات أو كيفيَّة تجهيز الوثائق والكلمة في محرك البحث. الجزء القادم يشرح عملية الفهرسة وعملية البحث بشكل تفصيلي.



الشَّكل ١-١: عملية استرجاع المعلومات.

٢- عملية الفهرسة

من أجل إتمام عملية البحث بشكل سريع ودقيق، يتم بناء فهرس لمجموعة الوثائق حتى تسهُل معرفة الوثائق التي تحتوي على مصطلحات معينة بشكل سريع.

٢,١- تحديد عنصر المصطلح

قبل عملية الفهرسة، يتم تجهيز المستندات والكلمات ووضعها في شكل مناسب من أجل إنشاء فهرس فعال في عملية البحث. من أهم العمليات عند تجهيز البيانات للفهرسة «تحديد عنصر المصطلح».

التعريف البديهي للمصطلح هو الكلمة. ولكن في الحقيقة - وفي معظم اللغات - فإنَّ تحديد عنصر المصطلح على أنه الكلمة لا يعد أحسن الخيارات لعملية البحث. وهذا لأنه في الغالب يمكن لكلمات عدة أن تكون أشكالاً سطحية مختلفة لنفس المصطلح. وهذا يشمل إضافة بعض الحروف لساق المصطلحات للحصول على كلمة مختلفة مثل إضافة اللواحق في اللغة الإنجليزية وإضافة السوابق واللواحق في اللغة العربية؛ انظر الجدول (١-١). وبالإضافة إلى الطرق المختلفة لكتابة نفس المصطلح كالهمزات في العربية (احمد/أحمد) والحروف الكبيرة (capital) في اللغات اللاتينية (ahmed/AHMED)، سيكون من المتوقع أن يجد المستخدم وثائق مسترجعة من عملية البحث تحتوي على المصطلح الأساسي في كلمات البحث بصرف النظر عن الشكل السطحي للكلمة. ولهذا فإنه من الضروري جدًّا لأي محرك بحث فعال أن تتم معالجة الكلمات المستخدمة في نصوص الوثائق وأيضا في كلمات البحث ليحدث تطابق للأشكال السطحية المختلفة من الكلمات التي ترجع لنفس المصطلح.

من أهم عمليات المعالجة الأساسية في معظم اللغات عملية التجريد (التَّجذيع) (stemming)، وهي عملية تجريد الكلمات من أي سوابق أو لواحق للحصول على ساق الكلمة مجردًا.

ويُوَضِّح الجدول (١-١) بعض الأمثلة لأشكال سطحية مختلفة لبعض الكلمات العربية والإنجليزية، والتي تقوم عملية التجريد بتوحيد هذه الأشكال المختلفة إلى شكل موحد، هو ساق الكلمة، مما يؤدي إلى تطابق أحسن عند البحث.

أمثلة	أشكال سطحية مختلفة لنفس ساق الكلمة	الساق
العربيَّة	الكتب، والكتب، كالكتب، فالكتب، كتبه، كتبها، كتبهم، كتبهن، كتبهما، وكتّبي، لكتبكم	كتب
	يلعب، تلعب، يلعبون، لعب، لعبت، وسيلعب	لعب
الإنجليزيَّة	play, played, playing, plays	Play
	Calculate, calculating, calculated, calculation, calculates	calculate

الجدول ١-١: أمثلة لبعض الأشكال السطحية لنفس المصطلحات في اللغة العربية والإنجليزية.

تُبَيِّنُ الأمثلة الموضحة في الجدول (١-١) أهمية تجريد الكلمات من السوابق واللواحق للحصول على الساق ليمثل المصطلح الذي سيدخل عملية الفهرسة. تتم هذه العملية أيضًا لكلمات البحث لتتم عملية التطابق على مستوى الساق للكلمات مما يؤدي إلى استرجاع أشكال مختلفة من نفس الكلمة، وبالتالي يؤدي إلى قدرة أعلى على استرجاع المعلومات.

هناك طرق مختلفة لتطبيق عملية التجريد للكلمات من السوابق واللواحق. أسهل هذه الطرق هي التي تعتمد على حذف حروف معينة من أول أو آخر الكلمات، مثل حذف حروف: «ال» ، «و» ، «ف» ، «وال» من أول الكلمات بالنسبة للغة العربية. ولكن المشكلة الكبيرة لهذه الطريقة هي عدم القدرة على تمييز الحُروف، أهِيَ أصليَّةٌ في الكلمة أم هي مجرد سوابق. هذا يتضح في كلمات مثل «الله» و «وحيد» ، لأن الناتج عن عملية التجريد بهذه الطريقة سيكون «له» و «حيد» على الترتيب. لهذا السبب، فإن هناك طرق أكثر تعقيدًا ودقة لحذف السوابق واللواحق دون حدوث أخطاء كهذه. أشهر هذه الطرق تعتمد على الأساليب الإحصائية والنماذج اللغوية للحصول على تجريد دقيق للكلمات.

ومن عمليات المعالجة - التي غالبا ما تطبق في كثير من اللغات - توحيد طريقة كتابة بعض الكلمات كما ذكرنا سالفا. ويكون الموضوع بسيطًا في بعض اللغات كالإنجليزية، حيث يتمّ توحيد الحروف الكبيرة لتصير كلها صَغيرة (case lower) من أجل تسهيل عملية التطابق (مثل: ahmed /AHMED). في لغة أخرى تكون عملية توحيد أسلوب

الكتابة للكلمات أكثر تفصيلا كاللغة العربية كما هو موضح في الجدول التالي:

عملية المعالجة	مثال	السبب
حذف التشكيل من الكلمات إن وجد	مُؤْمِنُونَ ← المؤمنون	ليست مستخدمة في أكثر النصوص العربية، ونادرا ما يستخدمها المستخدم في البحث
حذف علامة توسعة الكلمات إن وُجِدَت	مؤمنـــون ← مؤمنون	
توحيد الهمزات (ء، ئ، ؤ ← ء)	مؤمنون ← مءمنون ابناءه، ابناؤه، ابنائه ← أبناءه	لاختلاف كتابة الهمزات في بعض الكلمات حسب موقعها الإعرابي. وتطبيق هذه المعالجة لن يضر الكلمات الأخرى
توحيد رسم الألف والهمزة (ا، أ، آ، إ ← ا)	أحمد ← احمد إسلام ← اسلام آخر ← اخر	لعدم الالتزام بطريقة ثابتة لكتابة هذه الحروف في النصوص العربية، ولا حتى في كلمات البحث
توحيد الياء (ي، ى← ي)	أخرى ← أخري	
توحيد التاء المربوطة والهاء (ة، ه ← ه)	كلمة ← كلمه معالجة ← معالجه	

الجدول ١-٢: عملية توحيد طريقة كتابة الكلمات في اللغة العربية [للبحث أو الفهرسة].

من الممكن أن تكون هناك طرق معالجة أخرى للكلمات للحصول على عنصر المصطلح قبل عملية الفهرسة، وهذا يختلف من لغة إلى أخرى وفقًا لخصائص اللغة وطبيعتها.

٢,٢- حذف الكلمات المستَبعَدَة (words stop)

بالإضافة إلى تحديد عنصر المصطلح قبل عملية الفهرسة، والَّذي يكون في ساق الكلمات - غالبًا، فإن عملية حذف الكلمات المستَبعَدَة من النصوص تعد من أكثر الأساليب انتشارًا كإحدى عمليات المعالجة قبل الفهرسة. الكلمات المستَبعَدَة هي الكلمات ذات الأهمية الضعيفة في عملية البحث، والتي ليست لها قيمة قوية في تحديد إن كانت الوثيقة ذات صلة بموضوع البحث أم لا.

الكلمات المستَبعَدَة هي الكلمات التي تكون متواجدة في معظم الوثائق في مجموعة البحث، ولهذا فهي لا تميز وثيقة عن أخرى. وتشملُ هذه الكلمات في الغالب حُرُوفَ الجرّ (في، من، على، إلى) والضمائر (هو، هي، هم، أنت، أنتم).

هذه الكلمات لا تضيف معنى قويًّا إلى موضوع البحث وتكون متواجدة في معظم الوثائق، وبالتالي فإن حذفها يعد من العمليات التي تساعد على تحسين أداء محرك البحث، كما أن حذفها يساعد على تصغير حجم الفهرس وبالتالي تسريع عملية البحث.

اللغة	الكلمات المستَبعَدَة
العربية	في، من، على، إلى، عن، هو، هي، هم، هن، هما، هذا، هذه، هؤلاء، كنت، كان، له، لها، التي، الذي، قد، و، أو، أي، إن، إنه، إنها، ذلك، تلك ...
الإنجليزية	it, he, she, I, you, they, the, them, their, his, her, this, these, those, is, are, am, was, were, has, had, have, on, in, from, to, for, or, and, our, your ...

الجدول ١-٣: بعض أمثلة الكلمات المستَبعَدَة في العربية والإنجليزية

٢,٣- الفهرس

بعد عمليات المعالجة للنصوص في الوثائق والحصول على المصطلحات التي ستتم الفهرسة لها، يقوم نظام استرجاع المعلومات ببناء الفهرس لهذه المصطلحات. الفهرس هو جدول يحتوي على المصطلحات وقائمة الوثائق التي ظهر فيها كل مصطلح بالإضافة إلى إمكانية وجود معلومات أخرى عن المصطلح في حالة نظم استرجاع المعلومات المتقدمة.

بافتراض وجود مجموعة من الوثائق بحيث ترقم الوثائق ١، ٢، ٣، ... فإنَّ أبسط أشكال الفهرس يكون كما هو موضح في الجدول (١-٤):

المصطلح	أرقام الوثائق التي ظهر فيها
أحمد	١، ٦، ١٤، ٥٦، ٦١، ١٠٢، ....
مؤمن	٢، ٦، ٥٦، ٦١، ٩٨، ٢١٢، ....
أبناء	٦، ٤٣، ٦١، ٩٨، ١٤٥، ....

الجدول ١-٤: مثال لفهرس بسيط يستخدم في عملية استرجاع المعلومات.

تُمَثِّلُ الأمثلة في الجدول (١-٤) مجموعةَ مصطلحات بعد عملية تجريد السوابق واللواحق، وبعد عملية توحيد الحروف. الأرقام المقابلة لكل مصطلح هي أرقام الوثائق التي ظهر بها هذا المصطلح.

في أكثر الفهارس الحالية في نظم استرجاع المعلومات، يتم حفظ معلومات إضافية عن المصطلح في كل وثيقة. المعلومة الشائع حفظها هي عدد المرات التي ظهر فيها المصطلح داخل الوثيقة. كمثال لهذا، بإضافة المعلومة الجديدة يمكن أن تكون المعلومات المحفوظة للمصطلح «أحمد» داخل الفهرس كالآتي: (١،٤)، (١،٦)، (١٤،٢)، (٥٦،١٠). بمعني أن المصطلح «أحمد» ظهر داخل وثيقة ١ أربع مرات، ووثيقة ٦ مرة واحدة، ووثيقة ١٤ مرتَين، ووثيقة ٥٦ عشر مرات. هذه المعلومة تساعد على ترتيب الوثائق المسترجعة كما سيتم توضيحه في الجزئية القادمة.

أمثلة أخرى لبعض المعلومات التي يمكن حفظها عن المصطلحات والوثائق داخل الفهرس، يمكن أن تكون كالآتي:

- عدد الوثائق التي يظهر فيها المصطلح، وهي معلومة مهمة جدا تُظهر أهمية المصطلح. يتم شرح أهمية هذه المعلومة لترتيب الوثائق المسترجعة في الجزئية القادمة.
- طول كل وثيقة. حيث هناك طرق للبحث تقوم بمعاملة كل وثيقة حسب طولها. هذه المعلومة مهمة في الحالات التي يكون فيها طول الوثائق في المجموعة مختلفًا بشكل كبير.
- أماكن ظهور الكلمة في النص، حيث يحفظ مع كل مصطلح ترتيب ظهوره من بداية الوثيقة، وهذا يساعد عند الاستعلام عن أكثر من كلمة في كلمات البحث على معرفة قرب هذه الكلمات من بعضها، بحيث تساعد أيضًا على ترتيب الوثائق المسترجعة بجعل تلك التي فيها كل كلمات البحث أقرب من بعضها تأخذ ترتيبا أعلى في قائمة النتائج.
- حجم الخط الذي يكتب به المصطلح، وهذا يكون في النصوص المتقدمة والمنظمة كصفحات الويب. فمن المنطقي إعطاء أهمية أكبر للمصطلحات التي تكتب بخط أكبر كالعناوين وغيرها.

وبعد بناء الفهرس، تكون مجموعة الوثـائق جـاهزة لعملية البحث من قِبَل المستخدمين.

٣- آليَّة البحث

٣,١- تجهيز كلمات البحث

عندما يقوم المستخدم بإدخال كلمات البحث على محرك البحث، تكون أولى الخطوات هي تجهيز هذه الكلمة بالصورة المناسبة من أجل بحث الفهرس.

غالبا ما تكون العمليات المعالجية في عملية التجهيز هي نفسها التي تحدث قبل عملية الفهرسة. فبفرض أن عمليات التجهيز هي التي تم ذكرها سابقا في عملية الفهرسة من تجريد الكلمات وحذف الكلمات المستَبعَدَة؛ فإذا كانت كلمات البحث هي «أحمد والمؤمنون من أبنائه»، فإن مصطلحات البحث بعد المعالجة ستكون «أحمد مؤمن أبناء» لتكون جاهزة للبحث في الفهرس عن الوثائق التي تحتوي على هذه المصطلحات.

٣,٢- بحث منطقي (Binary Search)

يُعَدُّ البحث المنطقي أبسط أنواع طرق استرجاع المعلومات، حيث يعتمد فقط على تواجد كلمات البحث داخل الوثيقة ليقوم باسترجاعها دون محاولة ترتيب النتائج. ففي المثال السابق، تكون الوثائق المسترجعة هي التي تحتوي على المصطلحات الثلاثة «أحمد مؤمن أبناء» كلها، والتي تظهر في الجدول (١-٤) بأنها هي الوثائق: ٦ و٦١. هذه الوثائق المسترجعة تظهر في قائمة غير مرتبة على أساس صلة الوثيقة بموضوع البحث. ولكن فقط تكون كل الوثائق المسترجعة هي التي تحتوي على الثلاثة مصطلحات مجتمعة.

هذا الأسلوب في البحث غير واسع الانتشار إلا في بعض تطبيقات البحث القانونية كالبحث عن براءات الاختراع أو البحث في الوثائق القانونية، حيث يكون المستخدمون للبحث هنا أفرادًا متخصصين يقومون ببناء كلمة البحث بشكل حرفي يعتمد على وجود المترادفات في كلمات البحث. فيمكن أن تكون كلمات البحث في المثال السابق بالشكل الآتي: «أحمد + (مؤمن|تقي|ملتزم) + (أبناء|أولاد)». فعلامة «+» تدل على ضرورة وجود المصطلحات مجتمعة، وعلامة «|» تدل على إمكانية وجود أي

من الكلمات التي بين الأقواس. فيصبح معني المثال السابق: تحرَّ عن كل الوثائق التي تحتوي على المصطلح «أحمد» بالإضافة إلى أيٍّ من المصطلحات «مؤمن»، «تقي»، أو «ملتزم»، بالإضافة إلى أيٍّ من المصطلحين «أبناء» أو «أولاد».

٣,٣- بحث إحصائي

البحث الإحصائي هو الأكثر شيوعا وانتشارا في معظم محركات البحث في الوقت الحالي. هذه الطريقة من البحث تعتمد على النماذج الإحصائية لاسترجاع الوثائق. هناك نماذج إحصائية مختلفة لاسترجاع المعلومات ولكنها كلها في النهاية تعمل على استرجاع الوثائق التي تحتوي على معلومات تؤهلها لتكون ذات صلة بكلمات البحث وتعمل على ترتيبها ليكون الأكثر صلة على قمة قائمة النتائج. على عكس البحث المنطقي الذي تكون فيه النتائج هي التي تحتوي فقط على مصطلحات البحث، فإن البحث الإحصائي يعتمد على إعطاء كل وثيقة تحتوي على أي من مصطلحات البحث قيمة معينة، تزداد هذه القيمة كلما ازدادت دلالات صلة الوثيقة بمصطلحات البحث.

يمكن توضيح بعض الأفكار الأساسية لعمل هذه النماذج الإحصائية كالآتي:

- إعطاء وزن ضعيف للمصطلحات التي تظهر في عدد أكبر من الوثائق، حيث إنها دلالة على أن هذه المصطلحات غير قادرة على التمييز بين الوثائق المختلفة. وهذا هو السبب الأساسي لحذف الكلمات المستَبعَدَة التي تظهر في معظم الوثائق وتكون تقريبًا عديمة القيمة بالنسبة للبحث.
- إعطاء قيمة أكبر للوثائق التي تظهر فيها مصطلحات البحث بعدد أكبر. فكلما كانت مصطلحات البحث متكررة بشكل أكبر داخل الوثيقة كلما كان هذا دليلاً على أن الوثيقة تتحدَّث عن هذه المصطلحات.
- الاعتماد على نسبة ظهور المصطلحات داخل الوثيقة بدلا من عدد مرات الظهور، كنوع من إعطاء فرص متكافئة للوثائق القصيرة أمام تلك الطويلة. فظهور مصطلح معين عشر مرات في وثيقة طولها صفحة واحدة يمكن أن يكون أكثر صلة لهذا المصطلح من وثيقة أخرى ظهر فيها المصطلح عشرين مرة ولكن طولها عشر صفحات.

- إعطاء قيمة أكبر للوثائق التي تظهر فيها مصطلحات البحث قريبة أكثر من بعضها. فهذا يعد دليلاً على دقة وقرب الموضوع في الوثيقة من موضوع البحث.

هناك طرق إضافية لتحسين مستوى استرجاع المعلومات، ومعظمها يعتمد على إحصاء البيانات داخل مجموعة الوثائق، وفي بعض الأحيان تعتمد أيضًا على إحصاء البيانات من المصطلحات التي يستخدمها مُستخدمو محرك البحث.

كمثال توضيحي للنماذج الإحصائية في استرجاع المعلومات، يمكن النظر للمثال السابق «أحمد مؤمن أبناء»؛ حيثُ تكون كل الوثائق في الجدول (١-٤) قابلة للاسترجاع، لأن واحدة من هذه الكلمات على الأقل قد ظهرت بداخلها. ويتم ترتيب هذه الوثائق في النتائج على حسب نسبة ظهور كل مصطلح داخل أيٍّ من هذه الوثائق وأهميته.

٣,٤- إثراء كلمات البحث

أحد الأساليب المتبعة في بعض نظم استرجاع المعلومات هو إضافة بعض المصطلحات إلى المصطلحات الأصلية التي أدخلها المستخدم. الهدف الأساسي من هذا الأسلوب هو زيادة احتمالية التطابق بين موضوع البحث والوثائق للحصول على نتائج أفضل. هناك طرق متعددة لكيفية إضافة المصطلحات، لعلَّ أكثرها شيوعا:

- الإثراء بذات الصلة (feedback relevance): في هذه الطريقة تُعرض نتائج البحث على المستخدم ثم يقوم المستخدم بتمييز ما يراه ذا صلة لما يبحث عنه، وبعدها يعيد البحث بنفس كلمات البحث. يقوم محرك البحث باستخراج بعض الكلمات من الوثائق التي ميَّزها المستخدم بأنها ذات صلة وإضافتها إلى كلمات البحث الأصلية ليتم استرجاع وثائق جديدة تحتوي على الكلمة المضافة بالإضافة إلى كلمات البحث الأصلية. كمثال واضح لهذا، إذا كانت كلمات البحث الأصلية هي «أحداث ٩/١١» فمن المتوقع أن تكون معظم الوثائق ذات الصلة التي يحددها المستخدم تحتوي على الكلمات: «الولايات المتحدة الأمريكية» ، «الهجمات» ، «الإرهابية» ، «برجي التجارة». يقوم محرك البحث باستخراج هذه الكلمات بشكل آليّ وإضافتها إلى كلمات البحث الأصلية للحصول على نتائج بحث أفضل في المرة التالية.

- الإثراء المستَعار/ الزَّائِف بذات الصلة (feedback relevance pseudo): بما أن معظم المستخدمين لمحركات البحث لا يفضلون إجراء عملية البحث على مرتين، أو على الأقل ينأون عن تحديد ما يرونه ذا صلة من أجل تنفيذ عملية الإثراء، فهذه الطريقة تعتمد على الإثراء الاصطناعي دون الاحتياج إلى تدخل المستخدم من الأساس؛ فهي تفترض أن الوثائق المسترجعة على قمة قائمة النتائج تكون ذات صلة. ولهذا فهي الطريقة الأكثر انتشارا لعملية إثراء كلمات البحث. مثال ذلك: اعتبار أن الخمس وثائق المسترجعة من عملية البحث بكلمات البحث الأصلية تكون ذات صلة، ومن ثم تقوم باستخراج كلمات إضافية منها وإضافتها إلى كلمات البحث والبحث مرة أخرى. ما يتم عرضه للمستخدم هو نتائج البحث الثانية مباشرة مع عدم إظهار النتائج الأصلية له.
- معجم المترادفات: وهو معجم أو قاموس يحتوي على المصطلحات وبعض المرادفات لها. يمكن أن يكون هذا المعجم مبنيًّا من خلال لغويين أو مبنيًّا بطريقة آليَّة . قبل عملية البحث تضاف هذه المرادفات لمصطلحات البحث الأصلية ليتم البحث في مجموعة الوثائق عن التي تحتوي على كلمات البحث أو مرادفاتها.

ودائمًا ما تكونُ نتائج هذه الطرق لإثراء كلمات البحث غير ثابتة؛ فأحيانا تساعد هذه الطُّرُقُ على تحسين النتائج وأحيانًا تؤدي إلى الإضرار بها. وهذه مشكلة معهودة في معظم التقنيات وخصوصا ما يتعلَّقُ منها باسترجاع المعلومات؛ وهي مشكلة الدقة مقابل الكم. فكلما زادت المرادفات في كلمات البحث كلما زادت احتمالية استرجاع نتائج ذات صلة إضافية، ولكن في نفس الوقت يمكن استرجاع وثائق ليست ذات صلة. ولهذا يجب عند تصميم محرك بحث بخاصية إثراء كلمات البحث مراعاة أن النتائج لن تكونَ دائمًا أحسنَ ما تكون. وعليه، فمن الأفضل أن تكون هذه الخاصة اختيارية، بحيث يستطيع المستخدم الاستفادة منها أو تركها.

٤ - تقييم البحث

٤,١ - كيفية بناء مجموعات اختبار لاسترجاع المعلومات

يجب تقييم أداء محرك البحث للتأكد من قدرته الفعالة على استرجاع المعلومات ولمعرفة نقاط ضعفه والقدرة على تحسينها. من أجل عملية التقييم، يمكن بناء مجموعة بيانات لاختبار محركات البحث بطريقة علمية وعملية في نفس الوقت. مجموعة الاختبار ينبغي أن تحتوي على ثلاثة عناصر أساسية: مجموعة الوثائق، ومجموعة موضوعات البحث، وتحديد الوثائق ذات الصلة.

عنصر مجموعة الوثائق يكون في الغالب هو نفسه الذي يعمل عليه محرك البحث. وإن كان لا يوجد مجموعة معينة للوثائق ويراد اختبار محرك بحث معين أو طريقة بحث معينة، فيجب تحضير مجموعة بحث ذات طابع مناسب لمحرك البحث، وينبغي أن يكون عدد الوثائق في هذه المجموعة مقاربًا للواقع، بحيث لا يقل عن عشرات أو مئات الآلاف.

بالنسبة لمجموعة موضوعات البحث، يتم تجهيز مجموعة من الموضوعات الاختبارية ليتم البحث عنها في مجموعة الوثائق، وعند كتابتها يفضل مراعاة بعض الشروط:

- أن تكون مناسبة لمجموعة البحث المختبرة من حيث الطابع وأحيانا الفترة الزمنية. فعندما تكون مجموعة البحث عبارة عن مقالات إخبارية لإحدى الجرائد في فترة من الفترات، فليس من المتوقع أن تكون موضوعات البحث عن مقالات علمية في مجال الكيمياء، كما أنه ليس من المتوقع أن تكون موضوعات البحث عن أخبار في فترة زمنية تلي فترة مجموعة الوثائق بخمس سنوات، فغالبا ما تكون الأحداث مغايرة والأشخاص جُددًا.
- أن يوضح مع كل موضوع التفاصيل لما يتم البحث عنه بالتحديد ونوع وثائق المستندات المتوقع أن تكون ذات صلة. هذا التفصيل مهم جدًّا، حيث تحتمل كلمات البحث - كما ذكرنا آنفًا - أن تأخذ معاني مختلفة، بالإضافة إلى أن تقييم المستخدمين لما كانَ ذا صلة يختلف من شخص لآخر، ولذا يفضل دائما التفصيل في شرح ما ينبغي اعتباره ذا صلة.

- ألا يقل عدد موضوعات البحث عن ٢٥ موضوعًا. هذا الرقم بالتحديد جاء عن طريق عدة أبحاث في مجال استرجاع المعلومات؛ إذ وُجِدَ أنَّ هذا العدد هو أقل عدد لتكون النتائج الناتجة عن التقييم معبرة فعلا عن قدرة نظام البحث. بالطبع كلما زاد عدد الموضوعات كلما كان أفضل، حيث إن الرقم المتعارف عليه في كثير من الأبحاث في محال استرجاع المعلومات هو ٥٠ موضوعًا اختباريًّا.

أحداث ١١/ ٩	كلمات البحث
أحداث الهجمات على برجي التجارة العالميين في الولايات المتحدة الأمريكية في ١١ سبتمبر. ماذا حدث ومن المسئول عنها.	شرح موضوع البحث
الوثائق ذات الصلة ينبغي أن تتحدث في الأساس عن هذه الحادثة وتفاصيلها أو على الأقل النقاط الأساسية للموضوع . الوثائق التي تتناول تداعيات الهجمات دون الخوض في تفاصيل الهجمات نفسها لا تعد ذات صلة.	سرد الموضوع

الجدول ١-٥: مثال لموضوع بحث اختباري بعناصره التفصيلية، يُمكن استخدامه في عمليَّة التَّقييم.

يُوَضِّحُ المثال المعروض في الجدول (١-٥) أهمية وجود التفاصيل، حيث يمكن لأي مستخدم في هذه الحالة تقييم أي وثيقة إن كانت ذات صلة أم لا. وهذا يفتح الحديث عن العنصر الثالث لمجموعة الاختبار، وهو تحديد الوثائق ذات الصلة لكل موضوع.

تحديد الوثائق ذات الصلة هو ثالث عنصر أساسي لاستكمال عناصر تقييم البحث. ينبغي تحديد الوثائق ذات الصلة لكل موضوع بحث حتى يمكن بعد ذلك اختبار أي نظام استرجاع معلومات على قدرته على استرجاع تلك الوثائق.

الطريقة المثالية لتحديد كل الوثائق ذات الصلة بموضوع ما تتمثل في مراجعة كل الوثائق التي في المجموعة حتى لا يتم إفلات أي موضوع. بالطبع هذه الطريقة إن كانت مثالية فإنها غير واقعية بالمرة. فمن المستحيل مراجعة عشرات الآلاف من الوثائق، بل وأحيانا عشرات الملايين منها، لتحديد ما كانَ ذا صلة. الطريقة الواقعية هنا تعتمد على تحديد ما كانَ ذا صلة عن طريق مراجعة الوثائق المسترجعة من محرك البحث فقط. ولكن أي محرك بحث هذا؟ أهو الذي يُراد اختباره؟ كيف يكون ما يُراد اختباره هو نفسه الذي سوف يستخدم في تحديد ما كانَ ذا صلة؟ في هذه الحالة ستكون كل النتائج منحازة لهذا المحرك البحثيّ. ولهذا يتم استخدام أسلوب «التجميع» لحل هذه المشكلة.

أسلوب التجميع يعمل كالآتي:

- يتم البحث بموضوعات البحث باستخدام أكثر من محرك بحث وحفظ قائمة النتائج لكل محرك.
- يتم البحث بأكثر من طريقة واحدة في محرك البحث الواحد؛ فيمكن كمثال تفعيل إثراء كلمة البحث بطرق مختلف. كما يمكن استخدام كلمات البحث بالإضافة إلى شرح الموضوع لموضوعات البحث الاختبارية (الجدول ١-٥). يتم حفظ قائمة النتائج في كل مرة.
- يتم تجميع قوائم النتائج كلها في قائمة واحدة طويلة بعد حذف النتائج المتكررة. فيمكن أن تؤخذ قائمة نتائج تحتوي على ٥٠ وثيقة لكل طريقة بحث. بفرض تجميع ٢٠ قائمة، فالعدد النهائي للوثائق المسترجعة يمكن أن يكون ٥٠٠ بعد التأكد من عدم تكرار أي وثيقة مسترجعة في القائمة المجمعة.
- ترتب الوثائق في القائمة المجمعة بشكل عشوائي حتى لا يعطي انطباعًا بأن الوثائق في أعلى القائمة تكون ذات احتمالية أعلى لتكون ذات صلة.
- تعرض القوائم المجمعة لموضوعات البحث الاختبارية على مستخدمين ليتم مراجعة كل الوثائق في القائمة وتحديد ما كانَ ذا صلة بموضوع البحث بناءً على تفصيل الموضوع (كما هو موضح في الجدول ١-٥).
- يتم حفظ تقييم الوثائق إن كانت ذات صلة أو لا بالموضوع لتستخدم لاحقا في لتقييم أي محرك بحث.

بالطبع هذه الطريقة لا تضمن تحديد كل الوثائق ذات الصلة، ولكنها على الأقل تضمن إلى حد كبير استرجاع عدد كافٍ من الوثائق ذات الصلة، والأهم من هذا عدم انحيازها إلى محرك بحث أو طريقة بحث معينة.

عند اختبار أي نظام استرجاع معلومات لاحقا، يتم البحث بموضوعات البحث الاختبارية ومقارنة الوثائق المسترجعة بتلك التي تم تحديدها لتكون ذات صلة وحساب كمية الوثائق ذات الصلة التي نجح نظام استرجاع المعلومات المختبر في استرجاعها.

٤ , ٢ – نسبة الدقة (precision) مقابل نسبة الاسترجاع (recall)

الغرض من عملية البحث هو العثور على الوثائق ذات الصلة، حيث إنَّ أداء محرك البحث يقاس بمؤشرين رئيسين، هما مؤشر الدقة ومؤشر نسبة الاسترجاع. ويقوم مؤشر الدقة بحساب نسبة المستندات والوثائق ذات الصلة الناتجة عن عملية البحث مقارنة بالعدد الإجمالي للمستندات والوثائق الناتجة عن عملية البحث؛ بينما يقوم مؤشر الاسترجاع بحساب نسبة الوثائق ذات الصلة المسترجعة من عملية البحث مقارنة بالعدد الإجمالي للوثائق ذات الصلة. ولتيسير ذلك يمكن القول إن مؤشر الدقة يشير إلى قدرة المحرك على استرجاع وثائق ذات صلة ولكن ليست ضمن مجموعة كبيرة من الوثائق الأخرى. أما نسبة الاسترجاع فتشير إلى مدى نجاح المحرك في استرجاع أكبر كم ممكن من الوثائق ذات الصلة من مجموعة الوثائق.

المعادلة ١ والمعادلة ٢ توضحان كيفية حساب كلٍّ من الدقة ونسبة الاسترجاع:

نسبة الدقة = (عدد الوثائق ذات الصلة المسترجعة)/ (مجموع الوثائق المسترجعة)...... (١)

نسبة الاسترجاع = (عدد الوثائق ذات الصلة المسترجعة)/ (مجموع الوثائق ذات الصلة)........ (٢)

ما يمكن استنتاجه من المعدلات أن قيمة نسبة الاسترجاع تزيد كلما زاد عدد الوثائق المسترجعة، فهذا يعطي احتمالية أكبر لاسترجاع وثائق ذات صلة، ولكن في نفس الوقت غالبا ما يؤدي هذا إلى انخفاض الدقة لأن احتمالية استرجاع وثائق ليست ذات صلة يزيد أيضًا مع زيادة عدد الوثائق المسترجعة.

كمثال لحساب كلٍّ من الدقة ونسبة الاسترجاع، نفرض أنه تم اختبار أحد محركات البحث بأحد الموضوعات التي حدد لها ٥٠ وثيقة ذات صلة. بفرض أن محرك البحث المختبر قام باسترجاع ١٠٠ وثيقة، فإنه يمكن حساب الدقة ونسبة الاسترجاع عند نقاط مختلفة في قائمة النتائج كما هو موضح في الجدول (١-٦):

نسبة الاسترجاع	الدقة	عدد الوثائق ذات الصلة	عدد الوثائق المسترجعة (طول قامة النتائج)
٦/ ٥٠ = ٠,١٢	٦/ ١٠ = ٠,٦	٦	١٠
١٠/ ٥٠ = ٠,٢	١٠/ ٢٠ = ٠,٥	١٠	٢٠
٢٠/ ٥٠ = ٠,٤	٢٠/ ٥٠ = ٠,٤	٢٠	٥٠
٣٥/ ٥٠ = ٠,٧	٣٥/ ١٠٠ = ٠,٣٥	٣٥	١٠٠

الجدول ١-٦: مثال يوضح كيفية حساب كل من مؤشر الدقة ومؤشر نسبة الاسترجاع عند نقاط مختلفة من قائمة النتائج لموضوع اختباري له ٥٠ وثيقة ذات صلة.

كما يتضح من الجدول، فإنه في الغالب تقل الدقة كلما زادت أعداد الوثائق المسترجعة وعلى العكس تزيد نسبة الاسترجاع كلما زادت هذه الأعداد. ولهذا فإنه في مجال استرجاع المعلومات ينبغي أن يتم التوازن بين الدقة ونسبة الاسترجاع.

٤,٣- متوسط الدقة (precision average mean)

متوسط الدقة (MAP –precision average mean) هو المقياس الأكثر انتشارا لتقييم نظم استرجاع المعلومات. فهو يقيس متوسط الدقة عند نقاط مختلفة في قائمة النتائج. وكما أشرنا آنفًا، فإن الدقة تحسب عند نقطة معينة في قائمة النتائج، أما متوسط الدقة فهو يحسب على أنه متوسط قيم الدقة عند النقاط في القائمة التي توجد فيها وثيقة ذات صلة. وبعبارةٍ أخرى، تُحسَب الدقة كلما وُجدت وثيقة ذات صلة، ثم يتم حساب المتوسط لكل القيم المحسوبة. كمثال لهذا، إذا افترضنا وجود ستّ وثائق ذات صلة في النتائج العشرة المسترجعة الأولى في المراكز: {١، ٢، ٤، ٦، ٩، ١٠}، فإن متوسط الدقة يحسب كالآتي:

الدقة عند هذا المركز	عدد الوثائق ذات الصلة التي عثر عليها إلى الآن	المركز الذي توجد فيه وثيقة ذات صلة في قائمة النتائج
١/ ١ = ١	١	١
٢/ ٢ = ١	٢	٢
٣/ ٤ = ٠,٧٥	٣	٤

الدقة عند هذا المركز	عدد الوثائق ذات الصلة التي عثر عليها إلى الآن	المركز الذي توجد فيه وثيقة ذات صلة في قائمة النتائج
٤/٦ = ٠,٦٦	٤	٦
٥/٩ = ٠,٥٥	٥	٩
٦/١٠ = ٠,٦	٦	١٠

متوسط الدقة في هذه الحالة =

(١+١+٠,٧٥+٠,٦٦+٠,٥٥+٠,٦)/٦ = ٠,٧٦.

ولكن – لدقَّة حساب متوسط الدقة – ينبغي حساب متوسط الدقة عند كل النقاط التي توجد فيها وثيقة ذات صلة إلى أن يتم العثور على كل الوثائق ذات الصلة المحددة في مجموعة الاختبار. وبما أنه أحيانا يمكن ألا يتم استرجاع كل هذه الوثائق، فإنه يعتبر أن تلك الوثائق الغير مسترجعة وجدت عند المركز اللا نهائي لتكون الدقة في هذه الحالة مساوية للصفر؛ وبهذا فإن متوسط الدقة يحسب كما هو موضح في المعادلة ٣.

متوسط الدقة = مجموع قيم الدقة عند كل وثيقة ذات صلة في قائمة البحث / مجموع الوثائق ذات الصلة..........(٣)

فإذا افترضنا في المثال السابق أن عدد الوثائق ذات الصلة هو ثمانية وأن ما تم استرجاعه هو ٦ فقط، تكون قيمة متوسط البحث هي:

(١+١+٠,٧٥+٠,٦٦+٠,٥٥+٠,٦+٠+٠)/٨ = ٠,٥٧.

ما يمكن استنباطه من طريقة حساب متوسط الدقة أنه يركز على إيجاد الوثائق ذات الصلة على قمة قائمة النتائج، إذ إن إيجاد وثائق ذات صلة في مركز متأخرة في القائمة لا يضيف الكثير إلى قيمة القياس. ولهذا فإن متوسط الدقة يعطي أفضلية للنظم التي تستطيع أن تسترجع وثائق ذات صلة مبكرًا، وإن لم تجد كل الوثائق ذات صلة.

٥- محركات بحث الشَّبكة العنكبوتيَّة

تعد محركات بحث الشَّبكة العنكبوتيَّة [الويب] أكثر أنواع محركات البحث استخداما، حيثُ يَتَنَوَّعُ مُستخدموها في جميع أنحاء العالم. وتعتمدُ الفكرة الأساسية لمحركات بحث الويب على نفس فكرة استرجاع المعلومات؛ ولكنها تختلف عن محركات البحث العادية في عدَّة أشياء، منها:

٥,١- **مجموعة الوثائق**: من أهم الطباع الخاصة جدا بمحركات بحث الويب تنوع أشكال الوثائق التي يتم البحث فيها. فإن البحث يشمل صفحات الويب والصور، والمرئيَّات، والمقالات العلمية، والأخبار، وغيرها. كما أن أحجام مجموعة الوثائق يصل إلى مليارات الوثائق. هذا يعطي طابعا خاصا لمحركات بحث الويب حيث ينبغي أن تكون قادرة على معالجة هذا العدد الهائل من الوثائق بمختلف أنواعها.

٥,٢- **تجميع الصفحات والبيانات من على الإنترنت**: بخلاف معظم محركات البحث التي تكون فيها مجموعة الوثائق متواجدة ليكون كل ما على محرك البحث هو تنظيمها وفهرستها، فإن محرك بحث الويب يكون عليه أن يجمع الصفحات التي يريد فهرستها أولا من على الإنترنت. ولهذا فإن عملية تجميع البيانات تُعدّ من أهم عناصر فعالية محركات بحث الويب؛ فلا فائدة من وجود محرك بحث قوي للويب إن كان لا يحفظ الصفحات التي سيبحث فيها من الأساس. وبعض محركات البحث للويب تمتلك أفضلية على أخرى ليس بسبب أفضليَّة نظام البحث؛ وإنما لأنَّ أحدهما يستطيع تجميع صفحة الويب بشكل أكثر فعالية.

٥,٣- **التحديث المستمر**: المحتوى على الشَّبكة العنكبوتيَّة محتوى ديناميكي غير ثابت ويحدث له تحديث بشكل مستمر. ولهذا فإن محركات بحث الويب ينبغي أن تقوم بتحديث الفهرس أولا بأول لهذا الكم الهائل من الصفحات بشكل متكرر أيضًا. فبعض محركات بحث الويب تقوم بعملية التحديث للفهرس لبعض الصفحات عدة مرة في الساعة الواحدة لتواكب التغير المستمر في المحتوى للحصول دائما على نتائج بحث مستحدثة.

٥,٤- طريقة البحث: طريقة استرجاع المعلومات لمحركات بحث الويب تكون أكثر تقدما بكثير من محركات البحث العادية. فهي لا تعتمد على تطابق المصطلحات فقط، وإنما تمتد لتشمل خصائص كثيرة جدا منها:

- أهمية الصفحة: فإنها من أهم الخصائص الواجب أخذها في الاعتبار عند البحث، فليست كل الصفحات على الإنترنت تكون بنفس ذات الأهمية حتى وإن كانت تحتوي على نفس المحتوى. كمثال لأهمية هذه الخاصية، عند البحث عن كلمة «برنامج مايكروسوفت وورد» فإن النتائج يمكن أن تحتوي على إحدى الصفحات التي تشرح كيفية استخدام هذا البرنامج، ويمكن أن تكون هذه المصطلحات ظهرت مرات عديدة داخل الصفحة؛ فأي محرك بحث عادي سيضع هذه النتيجة في المركز الأول في قائمة النتائج. أما بالنسبة لمحرك البحث، فإن صفحة موقع شركة مايكروسوفت الرسمية على الإنترنت أكثر أهمية من تلك الصفحة، ولهذا فإن الصفحة النَّصِّيَّة على موقع الشركة، حتى وإن لم تظهر بداخله كلمات البحث إلا مرة واحدة، فإنَّه من المفضل أن تكون هي النتيجة التي على رأس القائمة ثم تليها بعد ذلك النتائج ذات الصلة الأخرى.
- سجل الاستخدام: أي محرك بحث ويب ناجح يقوم بتسجيل ما يقوم به مستخدموه من عمليات بحث في سجل حتى يستفاد منه لاحقا في تحسين أداء المحرك. يتم حفظ بعض المعلومات في هذا السجل ككلمات البحث التي يبحث عنها المستخدمون والنتائج التي يختارونها لهذه الكلمات. فإذا وجد في السجل أن معظم المستخدمين يقومون دائما باختيار النتيجة الرابعة لأحد موضوعات البحث، فهذا دليل قوي على أن هذه النتيجة هي أفضل من سابقاتها، ومن ثم يقوم محرك البحث بإظهارها على قمة النتائج بدلا من المركز الرابع.
- مكان المستخدم: يمكن لمحرك البحث معرفة مكان المستخدم عن طريق عنوانه الذي يقوم منه بعملية البحث. هذه المعلومة تساعد على

تحسين النتائج خصوصا لكلمات البحث التي فيها التباس، كالمثال الذي استُخدم في بداية الفصل عن «النادي الأهلي»، حيث يمكن لمحرك البحث تحديد النادي المقصود عن طريق معرفة مكان مُستخدم محرك البحث.

٥,٥- **كيفية التقييم**: تقييم أداء محركات بحث الويب يختلف قليلا عن نظم استرجاع المعلومات العادية. الفرق الأساسي هو تعريف ما كانَ ذا صلة، فالوثائق لا تحدد على أنها ذات صلة أو لا، وإنما ما كانَ ذا صلة يأخذ تقييما متدرجا، بحيث تحدد الوثائق في النتائج على كونها إجابة: -مثالية، ممتازة، جيدة، مقبولة، سيئة- وتُستخدم قياسات أخرى للتقييم تعتمد في الأساس على تقييم قدرة محرك بحث الويب على استرجاع النتائج الأفضل أولا. وغالبا ما تُحسب القياسات على استرجاع عشرة وثائق على الأكثر حيث إن مستخدم الويب في الغالب لا يقوم بتفحص أكثر من عشرة نتائج بحث.

تَدُلُّ هذه النقاط على أن نظم الاسترجاع الخاصة بالويب تكون أكثر تعقيدا وتقدما من محركات البحث الأخرى. ومن الأمثلة الشهيرة على محركات بحث الويب: جوجل (Google)، بينج (Bing)، ياهو (Yahoo)، ياندكس (Yandex)، بايدو (Baidu).

٦- محركات البحث المكتبية

محركات بحث المكتبات تعد أيضا من أكثر أنواع نظم استرجاع المعلومات انتشارا. ليس بالضرورة أن تكون هذه المحركات داخل المكتبات فقط، ولكنها أيضًا تشمل محركات البحث الخاصة بالكتب عامة كمواقع الكتب على الإنترنت. فكلها ذات طابع متماثل وتحتاج إلى طرق معالجة متشابهة. مجموعة الوثائق في هذه الحالة تكون عبارة عن كتب غالبا ما تكون ذات أعمار مختلفة.

ما يميز استرجاع المعلومات للمكتبات هو وجود محتوى الكتب القديمة. بما أن الكتب القديمة التي ترجع إلى ما قبل منتصف القرن العشرين تكون متواجدة فقط في صورة كتب مطبوعة، فمن أجل تفعيل عملية البحث لا بد من تحويل هذه الكتب إلى كتب رقمية تخزن على الحاسوب حتى يستطيع المستخدم البحث في محتواها بسهولة.

الطريقة المثلى لتحويل محتوى الكتب إلى شكل رقمي هي إعادة كتابتها وإدخالها للحاسوب عن طريق أشخاص متخصصين، ولكن هذه العملية يعيبها البطء الشديد والتكلفة الباهظة جدا لإدخال الآلاف وأحيانا مئات الآلاف من الكتب.

الحل البديل لعملية تحويل الكتب إلى الشكل الرقمي هو استخدام نظام التعرف الضوئي على الحروف OCR (التفصيل لنظم التعرف الضوئي على الحروف يوجد في الباب الخامس عشر)، بحيث يتم تحويل محتوى الكتب آليًّا إلى شكل رقمي وحفظ النص على الحاسوب لتفعيل القدرة على البحث. هذه الطريقة تتميز بالسرعة الفائقة والتكلفة الموفرة، ولكن تتمثل المشكلة الأساسية في وجود بعض الأخطاء في التعرف على بعض الحروف. وجود أخطاء في بعض الحروف يؤدي إلى عدم تطابق المصطلحات أثناء عملية البحث، وبالتالي يؤدي إلى انخفاض مستوى نتائج البحث. كمثال لهذه الأخطاء، إذا تم التعرف على كلمة «أحمد» في النص المطبوع على أنها «أحمر»، فهذا يؤدي إلى عدم استرجاع الوثيقة عند البحث عن كلمة «أحمد»، كما يؤدي إلى الاسترجاع الخاطئ لهذه الوثيقة عند البحث عن كلمة «أحمر».

هناك عدة أساليب متبعة من أجل تفادي هذه المشكلة الناجمة عن التعرف الخاطئ لبعض الحروف. يمكن ذكر أهمها كالآتي:

٦,١- **طريقة مطابقة المصطلحات**: في هذه الحالة تتم فهرسة المصطلحات بطريقة تؤدي إلى إمكانية التطابق النسبي بين الكلمات حتى في حالة وجود بعض الأخطاء. يكون تعريف المصطلح في هذه الحالة هو الشكل المتسلسل لحروف الكلمة، بحيث يتم استعراض الكلمة عن طريق متسلسلات الحروف الثنائية أو الثلاثية أو الرباعية للكلمة. كمثال، عند استعراض كلمة «أحمد» بالتسلسل الثلاثي للحروف تصبح كالآتي: «#أح أحم حمد مد#» بحيث يتم استعراض كل ثلاثة حروف متجاورة للكلمة وعلامة الشباك «#» تكون لتحديد بداية ونهاية الكلمات. في هذه الحالة، وعند التعرف الخاطئ على حرف الدال على أنه راء، تكون الكلمة محفوظة في الفهرس كالآتي: «#أح أحم حمر مر#»؛ فعندما يتم البحث عن كلمة «أحمد»، يتم تجهيز كلمة البحث بنفس الطريقة، فيكون التطابق بين الكلمة الصحيحة من

المستخدم والكلمة التي تحتوي على خطأ في الفهرس ٥٠٪ بدلا من صفر، حيث إن الكلمتين تشتركان في «#أح أحم». هذه الطريقة أثبتت فعاليتها في كثير من الأبحاث في هذا الموضوع للغات مختلفة، حيث إنها تؤدي إلى تحسين نتائج البحث للكتب المتعرف على نصوصها ضوئيًّا بشكل واضح.

٦,٢- **تصحيح الأخطاء في النصوص**: وهذه طريقة أخرى لتحسين نتائج البحث، حيث يتم استخدام نماذج اللغة وبعض الطرق الإحصائية لتصحيح هذه الأخطاء قبل عملية الفهرسة.

٦,٣- **إدخال الأخطاء على كلمة البحث**: هذه طريقة بسيطة يتم فيها استخدام بعض المعلومات من الإحصاءات عن طبيعة الأخطاء التي يمكن أن تحدث في التعرف الضوئي على الحروف، ثم يتم تطبيقها على كلمة البحث التي يُدخلها المستخدم بحيث تحتوي على كل احتمالات التعرف الخاطئ على كلمة البحث في الكتب. يتم التعامل مع هذه الكلمات على أنها مترادفات لتحسين عملية البحث. كمثال لهذا، عند إدخال كلمة «أحمد» في البحث، يمكن توقع أن تكون هذه الكلمة تم التعرف عليها خطأ في نصوص الكتب من بعد الإحصاءات على أنها: «أحمر» ، «أخمد» ، «أخمر» ...، فيتم اعتبار كل هذه الاحتمالات لكلمة «أحمد» على أنها مترادفات ليتم البحث عن أيها في الوثائق. هذه الطريقة أيضًا أثبتت فعاليتها في تحسين نتائج البحث في كثير من الأحيان.

هناك طرق معالجة أخرى لهذه المشكلة في محركات بحث المكتبة أو الكتب، كلها تعتمد على محاولة تفادي الأخطاء التي تحدث في عملية التعرف الآليّ على الحروف.

٧- محركات بحث شبكات التواصل الاجتماعي

مع التطور المطَّرِد للإنترنت وظهور ما يعرف بمواقع التواصل الاجتماعي كموقع فيسبوك (Facebook) وموقع تويتر (Twitter)، بات من الضروري وجود محركات بحث فعالة لتُمَكِّنَ المستخدمَ من الوصول إلى ما يحتاجه من معلومات على تلك المواقع.

تتميز مواقع التواصل الاجتماعي بعدة خصائص فريدة عن غيرها، تجعل عملية استرجاع المعلومات تواجه بعض التحديات. وتتمثَّلُ هذه الخصائص فيما يلي:

١ - **تنوُّع المحتوى**: يتميَّز محتوى المشاركات على شبكات التواصل الاجتماعي بالتنوع الكبير ما بين مشاركات نَصّية وصور ومرئيَّات وروابط خارجية. ويُضِيفُ هذا التنوع تحدِّيًا آخر في عملية استرجاع المعلومات من مواقع التواصل الاجتماعي.

٢ - **الكميات الكبيرة من المشاركات**: وصلَ عدد المشتركين في موقع تويتر في منتصف عام ٢٠١٢ إلى نصف مليار مشترك، يقومون بإرسال ما يزيد على ٢٠٠ مليون رسالة قصيرة يوميًّا عبرَ الموقع. وبالنسبة لموقع فيسبوك، فقد تخطَّى عدد المشتركين المليار مشترك في أوائل عام ٢٠١٣؛ ويتمُّ - كُلَّ ٢٠ دقيقة - وضع ما يزيد على مليون مشاركة على الموقع وإرسال أكثر من ٣ مليون رسالة خاصة. ويجعلُ هذا الكمُّ الهائل من المشاركات على مواقع التواصل الاجتماعيّ عمليةَ استرجاع المعلومات في غاية الصعوبة؛ بل يجعل حتى في عملية عرض نتائج البحث نفسها شيئًا من التحدي لكثرة وتنوع المحتوى.

٣- **اللغة المستخدمة**: وهي التي تميلُ في الغالب إلى العامّية. ويُعَبِّرُ معظم مُستخدمي مواقع التواصل الاجتماعي عمّا بداخلهم فيما يكتبونه، مما يجعلهم في معظم الأحيان يعبرون عنه بلهجة التخاطب العادية دونَ اللُّغة الرسمية. تتضح هذه الظاهرة بقوة في اللغة العربية بشكل خاص بسبب تعدد لهجاتها في مختلف الأقطار العربية. فهناك اللهجة المصرية والشامية والخليجية والمغربية وغيرها، وان كانت تجمع كلَّ هذه اللهجات لغةٌ رسمية واحدة. ولكن يوجد فوارق كبيرة بينها عند الاستخدام في مواقع التواصل الاجتماعي. ويُوَضِّحُ المثال في الجدول أسفله مثالاً على تنوع اللهجات في اللغة العربية:

اللهجة	الجملة
العربية الفصحى	ماذا تريد؟
المصرية	عايز ايه؟
الشامية	شو بدّك؟
الخليجية	ايش تبّي؟
المغربية	ويش تحب؟

يوضح المثالُ الاختلافَ الكبير بين مختلف اللهجات العربية في التعبير عن نفس المعنى. ويُوجِدُ هذا الاختلاف الكثير من التحديات أمام نظم استرجاع المعلومات، حيث تحتاجُ كلُّ لهجة من هذه اللهجات إلى عمليات معالجة خاصة بها. فطرق إضافة السوابق واللواحق في اللهجات العامية مختلفة عنها في اللغة الفصحى، كمثال («لم ألعب» في الفصحى، حيثُ تذهب في العاميات المختلفة إلى: «مالعبتش»، «مالعبت»، «مو لعبت»). وكذلك بالنسبة لمجموعة الكلمات المستبعدة الخاصة في العامّية، كمثال (اللي، ده، بس، مش، مو، عشان، ليه، ليش، ...).

وبما أن مواقع التواصل الاجتماعيّ نفسها لم تظهر إلا في السنوات الأخيرة، فإن الحلول البحثية لتحديات عملية استرجاع المعلومات لهذه المواقع ما زالت في خطواتها الأولى. ويستطيع المستخدم العادي الشعور بمشكلة البحث بنفسه على هذه المواقع، مثل: فيسبوك وتويتر. حيث يكون الوصولُ إلى معلومة معينة في منتهى الصعوبة. كذلكَ فإنَّ عرضَ النتائج لا يؤدي إلى الوصول للمطلوب بالشكل المرضي للمستخدم.

وعلى الرغم من هذا، فإن هناك العديد من الأبحاث لتحسين انطباع المستخدمين عن عمليات البحث على مواقع التواصل الاجتماعي. ويمكن تلخيص مجالات الأبحاث في استرجاع المعلومات من مواقع التواصل الاجتماعي في النقاط التالية:

٧,١ – دراسة دوافع البحث على هذه المواقع

كانت محاولة فهم دوافع المستخدمين للقيام بعمليات البحث على شبكات التواصل الاجتماعي من أقدم الدراسات للباحثين في مجال استرجاع المعلومات، والموضوعات التي يبحثون عليها وكيفيَّة مُقارنتها بالبحث على الويب. أظهرت تلك الدراسات أن دوافع البحث تكون في أغلب الأحيان لمعرفة آخر التحديثات والأخبار عن شخص أو حدث ما. وأكدت معظم الدراسات أن التحدي الأساسي في استرجاع المعلومات من هذه الشبكات يكون بسبب قِصَر المشاركة ولغتها. فالمشاركات تحتوي على عدد محدود من الكلمات بلغات دارجة وليست رسمية، فيكون العثور عليها صعبًا. وقد مهَّدَت هذه الدراسات الطريقَ لفهم عمليات البحث بشكل أحسن، كما حفَّزَت لبناء نظم استرجاع معلومات متخصصة لتلك البيانات.

٧,٢- عرض النتائج بشكل منظم

عند البحث عن موضوعاتٍ عامَّة في مواقع التواصل الاجتماعي، فإن النتائج تكون كثيرة جدًّا ومتنوعة. كمثال، عند البحث على تويتر باستخدام [هاشتاج (#tag)] لمتابعة آخر المشاركات عن موضوع معيَّن، تكون النتائج أحيانا بالآلاف، مما يجعل متابعة المشاركات المتعلقة عمليَّة صعبة. بالإضافة إلى أن هذه المشاركات تنقسم إلى آراء نصية وأخبار وروابط ومرئيَّات وصور وغيرها. أدى ذلك إلى استحداث وتطوير بعض النظم الخاصة لعرض النتائج بشكل منظم ومختصر للمستخدمين ليتسنى لهم معرفة المعلومات عمَّا يبحثون عنه. ويُعَدُّ «تويت موجز» (TweetMogaz) أحد هذه الأمثلة للمواقع المتخصصة في البحث في المشاركات (التغريدات) العربية على موقع تويتر. ولكنَّ طريقة البحث وعرض النتائج تختلف كليةً عن البحث على موقع تويتر نفسه. مبدئِيًّا، يتم البحث عن طريق تحديد كلمات البحث؛ وفي نفس الوقت يتم تحديد المدة الزمنية لاسترجاع التغريدات ذات الصلة ف خلالها؛ ثم يأتي الفارق الأساسي (في طريقة عرض النتائج)، حيث يتم معالجة كل المشاركات المسترجعة لاستخراج المشاركات الأكثر انتشارًا في الفترة الزمنية المحددة، وأيضا المشاركات الفكاهية، والمرئيَّات والصور الأكثر تداولا عبر المشاركات، والأخبار والمقالات التي يهتم بها المستخدمون عبر مشاركاتهم.

وتعطي هذه الطريقةُ المستخدمَ صورةً كلية عمَّا ينشره مُستخدمو المواقع الاجتماعية عن موضوع البحث؛ وهذا يعطي فكرةً عامَّة عن الرأي العام بالنسبة لموضوع معيَّن.

٧,٣- متابعة موضوعات بحث (Filtering Information)

وهو من أكبر تطبيقات علم استرجاع المعلومات، حيث يكون موضوعُ البحث ثابتًا. ويكون دور نظام استرجاع المعلومات هو تصنيف الوثائق والمستندات إلى ذات صلة أو غير ذات صلة بدلا من الترتيب. ويُستَخدَمُ هذا التطبيق في مجالات كثيرة، من أهمها: مُتابعة موضوعات البحث على شبكات التواصل الاجتماعي.

ومن أمثلة ذلك: قيام المستخدم بتحديد موضوع بحث عن شخص أو حادثة معيَّنة، ثم يقوم نظام البحث بتصنيف المشاركات الجديدة التي تظهر على أنها ذات صلة أم لا،

ثم يتم عرض المشاركات ذات الصلة أوّلاً بأوّل للمستخدم في حين ظهورها، وبهذا يكون متابعا للموضوع المتحرَّى عنه. وتتم عملية التصنيف بشكل آليّ بناءً على نموذج تصنيف مبنيٍّ من بعض الأمثلة الإيجابية والسَّلبيَّة للمشاركات ذات الصلة بموضُوعات مختلفة.

٧,٤- التنبؤ بالكوارث

يعد التنبؤ بالكوارث من أهم التَّطبيقات التي يتم دراستها في استرجاع المعلومات من شبكات التواصل الاجتماعيّ. بدأ هذا الموضوع يأخذ اهتماما كبيرًا بعد عام ٢٠١٠، حيث حدثت مُستَجَدَّاتٌ كبيرة حول العالم، كان لمواقع التواصل الاجتماعي تويتر وفيسبوك دورٌ كبيرٌ بها. من أمثلة ذلك: زلزال هاييتي ٢٠١٠، وبركان إندونيسيا ٢٠١٠، وإعصار ساندي في أمريكا ٢٠١٢، والثورات العربية منذ ٢٠١١، والاحتجاجات في اليونان وإسبانيا ٢٠١١.

لقد اكتشف الباحثون الدورَ الخطير لمواقع التواصل الاجتماعي، والتي تتحول وقتَ الأزمات بشكل خاص إلى مكان للاستغاثة والتنظيم ونقل الأخبار بشكل تعجز عنه وسائل الإعلام العادية. كل هذا دفع الكثيرين من الباحثين إلى عمل دراسات لمعرفة كيفية التنبؤ بالكوارث والأزمات عن طريق متابعة هذه المواقع وما يكتب عليها، بحيث تكون سببًا للتجهيز المسبق لتفادي الخسائر. بدأت هذه الأبحاث تُدعَم من قِبَل المنظمات الدولية، كالبنك الدولي والأمم المتحدة(1)، للوصول إلى طرق تلقائية لقياس أشياء اجتماعية لمناطق العالم المختلفة من هذه المواقع، مثل قياس مستويات الفقر والمرض والبطالة، بحيث تصل المساعدات الدولية إلى مستحقيها.

إن تقنيات استرجاع المعلومات لشبكات التواصل الاجتماعي لا تزال في بداياتها، والكثير من التطوير مطلوب لمواكبة الزيادة المطَّردة لهذه الشبكات التي لا يختلف اثنان على أهميتها في الحياة اليومية لمعظم مستخدمي الشَّبكة العنكبوتيَّة.

1- http://europeandcis.undp.org/blog/2013/01/11/can-big-data-help-deliver-better-operational-results/

٨- البحث الدلاليّ (Semantic Search)

يُمَثِّلُ البحثُ الدلاليُّ أحدَ خوارزمات البحث التي تأخذ في الاعتبار معاني الكلمات والمعنى السياقيّ للمصطلحات؛ وليس فقط النَّمط المماثل للحروف. وعلى الرغم من النجاح الحالي لمحركات بحث الويب الموجودة الآن إلا أن عمليَّةَ استرجاع المعلومات بصورتها الحالية ما زالت تُعانى قصورًا يتمثَّلُ في غياب فهم كلمات البحث ومعناها في السياق. ومن المتوقع أن يقوم البحث الدلاليّ بتعويض هذا النقص من خلال استخدام خوارزمات البحث التي تأخذ في الاعتبار معاني الكلمات والمعنى السياقيّ للمصطلحات مما يبشر بفرصة أكبر لزيادة دقة نتائج البحث والحصول على المزيد من النتائج ذات الصلة.

فمن خلال فهم معنى كلمات البحث وفهم معنى الكلمات الموجودة في مصادر البحث، يُتَوَقَّع أن تكون النتائج التي تَنتج عن عملية البحث متصلة بصورة أكبر بكلمات البحث وأن المصادر التي لم يكن في الإمكان الحصول عليها في نتائج البحث لعدم احتوائها بصورة مباشرة على كلمات البحث - بالرغم من أنها ذات علاقة بها - سوف تظهر في المعلومات التي تم استرجاعها.

ونظرا لما يبشر به البحث الدلاليّ من ثورة في مجال استرجاع المعلومات فقد قامت الشركات المنتجة لمحركات بحث الويب ذات الشهرة الواسعة، مثل: «جوجل» و «ياهو» و«بينج» .. باتخاذ الخطوات اللازمة نحو الاتجاه إلى هذه التقنية.

٨, ١- أمثلة للبحث الدلاليّ

إذا كان هناك محرك بحث يستخدم خوارزمات البحث الدلاليّ فإن إدخال سؤال مثل «من هي زوجة لويس الرابع عشر» في صندوق البحث لهذا المحرك سوف ينتج عنه أن يقوم هذا المحرك بعرض نتائج تتعلق بـ «مارى أنطوانيت» في النتائج ذات الصلة. وهذا دليل على أن هذا المحرك يستخدم البحث الدلاليّ وأنه قد قام بتحليل كلمات البحث وتبين له أن المستخدم يريد استرجاع معلومات عن زوجة لويس الرابع عشر وليس لويس الرابع عشر نفسه.

كذلك، عند قيام المستَخدِم بإدخال عبارة «هواتف خلوية» في صندوق البحث فسوف يقوم المحرك بعرض نتائج تحتوى على عبارات لها نفس المعنى والدلالة مثل «الهواتف النقالة» و «الموبايل» و «المحمول».

٨,٢- كيفية عمل محرك البحث الدلاليّ

هناك طريقتان تُستخدمان في عمل محرك البحث الدلاليّ:

- الترميز: وهو العنونة الدلالية للوثائق والكلمات والوحدات النصية الموجـــودة عـــلى صفحات الويب باستخدام الأنطولوجيا وإحدى لغات الويب الدلاليّ، مثل (OWL، RDFS، RDF، XML). ولا يتم عرض هذا الترميز لمتصفح الويب ولكن يمكن لمحرك البحث أن يستخدمه أثناء عملية الفهرسة بغرض الاستفادة من هذه المعلومات عند إجراء عملية البحث الدلاليّ.
- استخدام الذكاء الاصطناعيّ في فهم المعنى من السياق: فمثلا إذا رأى محرك البحث في صفحة على الويب أن ماري أنطوانيت هي زوجة لويس الرابع عشر فإنه يستنتج أن لويس الرابع عشر هو زوج ماري أنطوانيت. ويكون هذا بمثابة علاقة بين كلمتي البحث يمكن الاستفادة منها عند تكوين الفهرس؛ وبالتالي تتحقَّقُ الاستفادة عند البحث في هذا الفهرس.

٨,٣- تطبيقات البحث الدلاليّ في اللغة الإنجليزية

حيث إن البحث الدلاليّ قد عُنِيَ بتغيير الطريقة التي يتم بها البحث إلى الأحسن، لذلك فإن كثيرًا من المجهودات قد بُذلت بغرض إنتاج عدد من التطبيقات والأنظمة.

ويُعتبر (Wei et al 2008) مرجعًا جيِّدًا لبعض هذه الأنظمة؛ كما يُعَدُّ (SHOE) (Heflin & Hendler ٢٠٠٠ واحدًا من أقدم محركات البحث الدلاليّ؛ ويسمح للمستخدمين ببناء تساؤل منطقيّ عن طريق الأنطولوجيات. وبذلك يتطلب هذا النظام أن تكون المصادر التي يتم البحث فيها قد تم ترميزها / عنونتها دلاليًّا مسبقًا.

ومن أمثلة مُحرِّكات البحث الدلاليّ - كذلك:

- (SHOE): يُعَدُّ واحدًا من أقدم محركات البحث الدلاليّ؛ ويسمح للمستخدمين ببناء تساؤل منطقيّ عن طريق الأنطولوجيات. وبذلك يتطلب هذا النظام أن تكون المصادر التي يتم البحث فيها قد تم ترميزها / عنونتها دلاليًّا مسبقًا.
- (KIM, OWLIR): وتم الاعتماد فيهما على استخدام الاستدلال المنطقيّ وطرق استرجاع المعلومات التقليدية معا. ففي حالة ما لم يتم الحصول على نتائج باستخدام البحث الدلاليّ يتحول النظام إلى الطريقة التقليدية في استرجاع المعلومات.
- (Aqualog): وهو نظام للإجابة عن الأسئلة باستخدام الدلالات.

هذا بالإضافة إلى محركات بحث الويب المذكورة آنفًا، مثل: «جوجل» و «ياهو» و«بنج»؛ والتي اتجهت بالفعل إلى استخدام خوارزمات البحث الدلاليّ.

٨ , ٤ - تطبيقات البحث الدلاليّ في اللغة العربية

لا تزالُ الأبحاثُ المعنِيَّة باسترجاع المعلومات العربيَّة محدودةً إلى درجةٍ كبيرة. ومنها على سبيل المثال:

- (El-Beltagy et al 2003). قام البحث باستغراق إضافة بيانات تكميلية إلى قصاصات (Snippets) المعلومات الزراعية في إحدى التجارب واستخدامها لتحسين استرجاع القصاصات التي لها صلة بكلمات بحث المستخدم.
- (Zaidi and Laskri 2005). في هذا العمل تم استخدام أنطولوجيا خاصَّة بالحقل القضائيّ (Legal domain) مع آلية استرجاع المعلومات.
- (Qawaqneh, 2007). يقدم طريقة لترتيب النتائج باستخدام مبدأ الأنطولوجيا. وتقوم الطريقة المقترحة بترتيب الوثائق بالاعتماد على عدد مرات تكرار مبادئ الأنطولوجيا التي تظهر في الوثائق.
- (Semahtic MediaWiki). قامت الدِّراسة بإضافة اللغة العربية إلى قائمة اللغات التي يمكن أن تتعامل مع الترميز الدلاليّ عند إنشاء صفحات Wiki الحُرَّة لها. وبالتالي أتاحت للناشرين أن يقوموا بنشر محتوى ويب دلاليّ.

- ومن ناحية أخرى، أعلَنَت بعضُ الشركات التي تعمل في مجال البحث على الويب أنها سوف تقوم بتقديم محركات بحث دلاليّ للغة العربية. ومن أمثلة ذلك: (Kngine) و(The next web).

٩- أفكار تصلح للأطروحات العلميَّة (الماجستير والدكتوراه)

٩,١- تجهيز مجموعات اختبارية لاسترجاع المعلومات

إحدى الأفكار التي تصلح لأن تفرز رسائل ماجستير، هي إعداد مجموعات اختبارية للبحث. يمكن أن يقوم الباحث باختيار إحدى مجموعات الوثائق ذات الطابع المحدد والقيام بتجميع مجموعة الوثائق وترتيبها بشكل منظم. ينبغي مراعاة الشروط والخصائص التي تم توضيحها آنفًا في المجموعات الاختبارية، بحيث يكون عدد الوثائق مناسبًا لطبيعة المجموعة، فلا يقل عن عشرات الآلاف.

وينبغي أيضًا أن تكون موضوعات البحث الاختبارية مناسبة. ويفضل في حالة بناء المجموعة الاختبارية عن طريق فريق بحث واحد أن يقوم بالاستفادة من متطوعين لاختيار موضوعات البحث الاختبارية، وأيضًا لتحديد الوثائق ذات الصلة. لتفادي انحياز نتائج البحث إلى طريقة بحث واحدة، يفضل استخدام محركات بحث مختلفة، حيث يتوفر عددٌ منها مجانا من أجل الأغراض البحثية مثل: Indri، Lemur، Lucene، Terrier وغيرها من محركات البحث المجانية التي يستطيع الباحث أن يستخدمها من أجل فهرسة مجموعة الوثائق واستخدامها في البحث عن موضوعات البحث بنماذج وآليَّات بحث مختلفة لمحرك البحث الواحد. بهذا يمكن للباحث استخدام عملية تجميع النتائج بسهولة من أجل تحديد ما كانَ ذا صلة بطريقة علمية سليمة ودون انحياز.

بالنسبة لمجموعة الوثائق التي يمكن تجميعها وتجهيزها، يمكن أن تكون:

- صفحات ويكيبيديا: يمكن تحميل كل مقالات ويكيبيديا لأي من اللغات من على الموقع نفسه، ثم اختيار موضوعات البحث المناسبة لها وتحديد ما كانَ ذا صلة من المقالات بعد ذلك.

- كُتُب: في مجال معين أو في تخصصات مختلفة. يمكن أن تكون الكتب الإلكترونية في الأصل؛ ويمكن أن تكون من التي تم التعرف على محتواها آليًّا، ولكن في هذه الحالة تكون طُرُق المعالجة مختلفة كما أسلَفنا.
- مقالات علمية أو أطروحات علميَّة (ماجستير ودكتوراه): يمكن أن تكون مجموعات الوثائق ذات طابع علمي في مجال معين، ويتم اختيار موضوعات البحث بناءً على ذلك. ولكن في هذه الحالة ينبغي مراعاة أنَّ من يقوم بتحديد ما كانَ ذا صلة على علم بهذا المجال، أو على الأقل لديه بعض الخلفية عن المجال العلمي.
- مقالات إخبارية: من نفس المصدر كمقالات إحدى الجرائد لأعوام متعددة أو من مصادر إخبارية مختلفة. يمكن أن تكون المقالات من مجال إخباري معين، كالرياضة أو السياسة أو الفنون أو غيرها. المهم في أي حالة هو اختيار موضوعات البحث الاختبارية بما يناسب طبيعة مجموعة الوثائق.
- مجموعات من الصور أو المرئيَّات: تتمثَّل الطريقة الأسهل في اختيار المجموعات التي تكون الصُّوَرُ أو المرئيَّات فيها مصحوبةً بمُسمى أو شرح لمحتوى هذه الصور والمرئيَّات، مثل الصور على موقع «فليكر Flickr» والمرئيَّات على موقع «يوتيوب Youtube».

٩,٢- استرجاع المعلومات من شبكات التواصل الاجتماعي

كما أوضحنا مسبقا، فإن مجال البحث في هذا الموضوع مازال في إرهاصاتِه. وهناك الكثير من الأفكار التي يمكن تطويرها لخدمة استرجاع المعلومات من مواقع التواصل الاجتماعي وللَّهجات الدّارجة بشكل عام.

و يمكن للأفكار البحثية المقتَرَحة من دراسة هذا الموضوع أن تكون أطروحات ماجستير أو دكتوراه في مجالات مختلفة.

ومن هذه الموضوعات:

- بناء مجموعة اختبارية لاسترجاع المعلومات للمواقع الاجتماعية: ينبغي أن تُراعى بشدة الطبيعة الخاصة بهذه البيانات في كيفية اختيار الموضوعات وكيفية كتابتها.

كما أن كيفية اختيار ما كانَ ذا صلة يتطلب جهدًا أكبر في معالجة البيانات لبناء محركات بحث مختلفة لتنفيذ عملية التجميع للنتائج أو بحث طرق أخرى من أجل تحديد ما كانَ ذا صلة من الوثائق (مشاركات المستخدمين في هذه الحالة).

- محاولة استنتاج طريقة معرفة أساليب الكتابة المختلفة للموضوعات الاجتماعية وتشكيل طريقة ممنهجة لتوحيد طرق الكتابة التي من المتوقع أن تتعدى توحيدَ بعض الحروف أو تجريد الكلمات؛ كما يمكن تحديد مجموعة جديدة من الكلمات المستَبعَدَة للهجات الدَّارجة.
- تصنيف المشاركات التي يتم البحث عنها بطرق مختلفة، مثل تصنيفها حسب الموضوع: سياسي، اجتماعي، ترفيهي ...، أو تصنيفها حسب حالة الكاتب: سعيد، حزين، غاضب ... وغيرها من التصنيفات. كل هذا يمكن أن يفيد لاحقًا في القدرة على استرجاع المعلومات.
- تجميع المشاركات الَّتي تتناول نفس الموضوع آليًّا. ويُعدُّ هذا مفيدًا للغاية كأحد الخصائص لهذه الموقع الاجتماعية، حيث سيكون من المفيد للمستخدم أن يجد كل المشاركات التي تتحدث عن نفس الموضوع مجمعة تلقائيًّا. إنَّ بناء نظام يقوم بهذا يمكن أن يكون أطروحة دكتوراه، ويمكن أيضًا عمل أطروحة ماجستير في تجهيز البيانات ومجموعة اختبار تساعد على بناء نظام كهذا.

٩,٣- الصفحات الشخصية

الصفحة الشخصية أو ما يعرف بالتَّدوينات الإلكترونيَّة (Blogs) هي صفحات خاصة بالمستخدمين، يَكتبُ كل شخص آراءه فيها على هيئة مقالات. وتُشبهُ اللغةُ المستخدمة في هذه التَّدوينات تلكَ التي تستخدم في الصفحة الاجتماعية، حيث يمكن أن تأخذَ أشكالاً مختلفة. يمكن أن تكون إحدى أفكار الماجستير أو الدكتوراه بناء مجموعة اختبارية لهذه الأشكال من الصفحات وتطوير طرق استرجاع فعالة لها.

٩,٤- استرجاع المعلومات عبر اللغات

من أهم الموضوعات البحثية في علم استرجاع المعلومات. والهدف هو كتابة موضوع البحث بلغة ما، وتكون المعلومات والوثائق المسترجعة من لغة أخرى.

الأبحاث في هذا الموضوع بالنسبة للغة العربية محدودة جدًّا وتركز على البحث بين اللغة العربية والإنجليزية. يمكن بناء مجموعات اختبارية لاختبار البحث عبر اللغات المختلفة مع اللغة العربية؛ ويمكن عمل هذا بشكل بسيط عند بناء أي مجموعة بحث عادية بالقيام بترجمة موضوعات البحث الاختبارية ترجمة يدوية إلى لغات أخرى لتكون مُعَدَّة لاختبار البحث عبر اللغات؛ كما يمكن بناء مجموعات اختبار مخصصة للبحث عبر اللغات، وهذا بالنسبة للمجموعات التي تحتوي على وثائق من لغات متعددة.

١٠- مِن المواقع الإلكترونيَّة التَّعليميَّة والإرشاديَّة

١- محركات بحث مجانية لغرض البحث العلمي:

- Indri, Lemur: http://www.lemurproject.org/
- Lucene: http://www.getopt.org/luke/
- Terrier: http://terrier.org/
- Solr: http://lucene.apache.org/solr/

٢- قوائم بالكلمات المستَبعَدَة للغات متعددة:

- http://members.unine.ch/jacques.savoy/clef/index.html

٣- أدوات تجريد الكلمة من السوابق واللواحق للغات مختلفة:

- http://snowball.tartarus.org/

٤- مواقع بحث لشبكات التواصل الاجتماعي:

- http://www.tweetmogaz.com
- http://www.topsy.com
- http://bottlenose.com

ببليوجرافيا مرجعيَّة

1. Abu El-Khair, I. (2007). Arabic Information Retrieval. Annual Review of Information Science and Technology, Vol. 41, Issue 1.
2. Agrawal ,A ;.Agrawal ,A .J .(2016) .Designing Cross-Language Information Retrieval System using various Techniques of Query Expansion and Indexing for Improved Performance. IRJET, Vol. 03, Issue 02.
3. Baeza-Yates ,J ;.Ribeiro-Neto ,B .(2010) .Modern Information Re-trieval :The Concepts and Technology behind Search. ACM Press Books.
4. Bansal ,M ;.Arora ,J .(2016) .A Novel OBIRS System for Ontology Based Information Retrieval System. IJEDR, Vol. 4, Issue 2.
5. Codocedo ,V ;.Lykourentzou ,I ;.Napoli ,A .(2014) .A semantic ap-proach to concept lattice-based information retrieval. Annals of Mathematics and Artificial Intelligence, Vol. 72, Issue 1, pp. 169-195.
6. Darwish ,K ;.Magdy ,W ;.Mourad ,A .(2012) .Language Processing for Arabic Microblog Retrieval. CIKM 2012.
7. Darwish ,K .(2002) .Douglas W .Oard :CLIR Experiments at Maryland for TREC :2002 Evidence Combination for Arabic-English Retrieval. TREC 2002.
8. Darwish ,K .(2002) .Douglas W .Oard :Term selection for searching printed Arabic. SIGIR 2002.
9. El-Beltagy ,S ,.R ;.Rafea ,A ;.Abdelhamid ,Y" .(2003) .Chapter:13 Using Dynamically Acquired Background Knowledge for Information Extraction and Intelligent Search ."In Editor) Masoud Mohammadian ,(Intelligent Agents for Data Mining and Information Retrieval ,Hershey ,PA ,USA.
10. Haynes ,D .(2018) .Metadata for Information Management and Retrieval :Understanding metadata and its use. Facet Publishing.

11. Latha, K. (2017). Experiment and Evaluation in Information Retrieval Models. CRC Press.
12. Lee, P.; West, J. D.; Howe, B. (2016). VizioMetrix: A Platform for Analyzing the Visual Information in Big Scholarly Data. ACM.
13. Liu ,X ;.Gao ,J ;.He ,X ;.Deng ,L ;.Duh ,K ;.Wang ,Y .(2015) .Representation Learning Using Multi-Task Deep Neural Networks for Semantic Classification and Information Retrieval. Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pp. 912-921.
14. Lupu ,M ;.Mayer ,K ;.Kando ,N ;.Trippe ,A .(2018) .Current Challenges in Patent Information Retrieval. Springer Berlin Heidelberg.
15. Magdy, W.;Darwish, K. (2007). Error correction vs. query garbling for Arabic OCR document retrieval. ACM Transaction of Information Systems, Vol. 26, Issue 1.
16. Magdy ,W ;.Darwish ,K .(2008) .Effect of OCR error correction on Arabic retrieval .Springer Information Retrieval, Vol. 11, Issue 5.
17. Magdy ,W ;.Darwish ,K .(2010) .Omni font OCR error correction with effect on retrieval. ISDA 2010.
18. Manning ,C .D ;.Raghavan ,P ;.Schütze ,H .(2009) .Introduction to Information Retrieval. CUP.
19. Naveed ,T ;.Gottron ,J ;.Kunegis ,A .(2011) .Searching microblogs: coping with sparsity and document quality. CIKM 2011.
20. Palangi ,H ;.Deng ,L ;.Shen ,Y ;.Gao ,J ;.He ,X ;.Chen ,J ;.Song, X ;.Ward ,R .(2016) .Deep Sentence Embedding Using Long Short-Term Memory Networks :Analysis and Application to Information Retrieval. IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 24, Issue 4.
21. Perez ,J .C" .(2009) .Google Rolls out Semantic Search Capabilities ,"http//:www.pcworld.com/businesscenter/article/161869/google_rolls_out_semantic_search_capabilities.html

22. Qawaqneh ,Z ;.El-Qawasmeh ,E ;.Kayed ,A" .(2007) .New Method for Ranking Arabic Web Sites Using Ontology Concepts ,"in proceedings of IEEE ICDIM ,p.649-656 .

23. Shaila ,S ;.Vadivel ,A .(2018) .Textual and Visual Information Retrieval using Query Refinement and Pattern Analysis. Springer Singapore.

24. Teevan, J.; Ramage, D.; Morris, M. R. (2011). #Twittersearch: A comparison of microblog search and web search. WSDM 2011.

25. Verma ,A ;.Kaur ,I ;.Singh ,I .(2016) .Comparative Analysis of Data Mining Tools and Techniques for Information Retrieval. Indian Jour-nal of Science and Technology, Vol. 9 (11).

26. Vicente-López, E.; de Campos, L. M.; Fernández-Luna, J. M.; Huete, J. F.; Tagua-Jiménez, A. & Tur-Vigil, C. (2015). An au-to-matic methodology to evaluate personalized information retrieval systems. User Modeling and User-Adapted Interaction, Vol. 25, Issue 1, pp. 1-37.

27. Wei ,W ;.Barnaghi ,P .M ;.Bargiela ,A" .(2008) .Search with Meanings :An Overview of Semantic Search Systems ,"International journal of Communications of SIWN ,Vol ,3 .pp.76-82 .

28. Zaidi ,S ;.Laskri ,M .(2005) .A cross-language information retrieval based on an Arabic ontology in the legal domain. The International Conference On Signal-Image Technology & Internet-Based Systems (SITIS'05), Morocco.

الفصل الثَّاني

التَّرجَمة الآليَّة

د. أحمد رافع

١- نظرة عامَّة مُوجَزة.

٢- تعريف بأهم المصطلحات المستَخدَمة في التَّرجمة الآلية.

٣- تقنيات التَّرجمة الآلية، وآخر التَّوجُّهات البحثِيَّة.

٤- البرامج والموارِد اللغوية المرتبطة بالتَّرجمة الآلية.

٥- أهم المواقع والأدوات المساعدة للموارِد والتِّقنيات مفتوحة المصدر.

٦- أفكارٌ لتطوير مُدَوَّنات لُغَوِيَّة مُستَقبليَّة لأهداف التَّرجمة الآليَّة.

ملحق - الأساس النَّظريّ لبناء نظام ترجمة آليّ إحصائيّ.

في عام ١٩٤٩ بدأت أبحاث التَّرجمة من الرُّوسية إلى الإنجليزية في أعقاب الحرب العالمية الثانية. وفي عام ١٩٥٤ تم استحداث أول نموذج لنظام الترجمة الآلية من الروسية إلى الإنجليزية. وبعد اثني عشر عاما من البحث صدرَ تقرير من اللجنة المشكلة من قبل الهيئات الحكومية المانحة لأبحاث الترجمة الآلية في الولايات المتحدة الأمريكية بعدم جدوى القيام بالأبحاث في هذا المجال. وقد كان أثر هذا التقرير سيئا على تقدم البحث والتطوير في مَيدان الترجمة الآلية. لذلك اتَّجهت الأبحاث في السبعينيَّات من القرن الماضي إلى كندا وأوروبا الغربية. وفي الثمانينيَّات بدأت تظهر تقنيات المنهج التحويلي وقواعد المعرفة؛ كما ظهرت تقنيات التَّرجمة القائمة على أسس إحصائِيَّة في التسعينيَّات. وقد استُخدِمت هذه التقنيات في الترجمة من العربية إلى الإنجليزية بكثرة في بداية الألفية الثانية؛ وذلكَ لاهتمام الولايات المتحدة بالترجمة من اللغة العربيَّة بعد أحداث سبتمبر ٢٠٠١.

يُقَدِّمُ هذا البابُ المفاهيم والمصطلحات الأساسية للترجمة الآلية، ويستعرض الطرق الرئيسة لها بصورة مبسطة مع إعطاء أمثلة توضح الفكرة العامة لكل طريقة. ويتكون البابُ من سبعة أقسام على النَّحو التَّالي:

١-يقدم القسم الأول نظرة عامة موجزة عن الطُّرُق الرَّئِيسة الثَّلاثة للتَّرجمة الآلية. وهي: طريقة قائمة على قواعد لغوية، وطريقة قائمة على استخدام أمثلة لجمل سبق ترجمتها، وطريقة قائمة على أسس إحصائِيَّة باستخدام كَمٍّ كبير من النُّصُوص المتَرجَمة.

٢-ويُعَرِّفُ القسم الثاني أهمَّ المصطلحات المستخدمة في كل تقنية من تقنيات التَّرجمة الآلية المشار إليها في القسم الأول.

٣-أمَّا القسم الثالث فيُلقي الضَّوءَ على تقنيات التَّرجمة الآلية، وآخر التَّوجُّهات البحثِيَّة؛ وخُصُوصًا المنهج القائم على مستوى ترجمة بناء هرمي للعبارة والمنهج القائم على مستوى ترجمة لغة المصدر إلى بناء نحوي للغة الهدف. كذلك يُعنى باستخدام تقنية التَّعلُّم العميق للتَّرجمة الآليَّة.

٤-ويَصِفُ القسم الرابع البرامج اللغوية المطلوبة لنظم التَّرجمة الآلية القائمة على القواعد اللغوية ونظم التَّرجمة الآلية القائمة على الأمثلة ونظم التَّرجمة الآلية القائمة على أسس إحصائِيَّة.

٥-أمَّا القسم الخامس فيستَعرضُ أهم المواقع والأدوات المساعدة للموارِد والتِّقنيات مفتوحة المصدر.

٦-ويَقترح القسم السادس أفكارًا لتطوير مُدَوَّناتٍ ثنائيَّة اللُّغة باستخدام المادة المترجمة المتاحة على شبكة الويب، مثل مواقع الأمم المتحدة. وكذلك إضافة قيمة للمدَوَّنات ثنائيَّة اللُّغة المتاحة لتحسين جودة نموذج الترجمة الناتج عن هذه المدَوَّنات.

٧-وأخيرًا، يُقَدِّمُ القسم السَّابع ببليوجرافيا مرجعيَّة، تشتملُ على مجموعةٍ من المراجع الَّتي تعرضُ للمفاهيم والمصطلحات الرئيسة في الترجمة والترجمة الآلية، ولتاريخ الترجمة الآلية، وكيفيَّة التعرُّف على الاستعارات والتعبيرات المجازية وكيفية ترجمتها؛ بالإضافة إلى تقنيات التَّرجمة الآلية، وآخر التَّوجُّهات البحثِيَّة، وبعض نظم الترجمة الآلية من وإلى اللغة العربية.

١ - نظرة عامَّة مُوجَزَة

يمكن تصنيف طُرُق التَّرجمة الآلية - عُمُومًا - إلى أربع طُرُقٍ رئيسية: طريقة قائمة على قواعد لغوية، وطريقة قائمة على استخدام أمثلة لجمل سبق ترجمتها، وطريقة قائمة على أسس إحصائِيَّة باستخدام كَمٍّ كبير من النُّصُوص المتَرجَمة، وطريقة قائمة على استخدام تقنية التعلم العميق للترجمة الآلية.

ويمكن تصنيف مناهج التَّرجمة الآلية القائمة على القواعد إلى: المنهج المباشر والمنهج التحويلي، ومنهج اللغة الوسيطة. والفرق بين هذه المناهج هو مقدار التحليل اللغوي الذي يتم عمله على لغة المصدر وتحويل ناتج هذا التحليل إلى مفردات وقواعد اللغة المستهدفة ثم توليد لغة الهدف باستخدام قواعد الصَّرف وبناء لغة الهدف.

تتميز الطريقة القائمة على استخدام الأمثلة للترجمة الآلية بأنها تستخدم مجموعة من النصوص المتَرجمة المتوازية، والتي تم تمثيل قاعدتها المعرفية الرئيسية؛ والأساس الذي تقوم عليه هذه الطريقة هو التَّرجمة عن طريق القياس، حيثُ يعتمد هذا الطريق على الاعتقاد بأن الناس تقوم بالتَّرجمة عن طريق تحليل الجملة إلى عبارات ثم ترجمة هذه العبارات وتجميعها في جملة واحدة.

طريقة التَّرجمة القائمة على أسس إحصائِيَّة تستخدم فيها النماذج الإحصائِيَّة معلماتها مشتقة من تحليل كم كبير من النصوص - ثنائية اللغة وأحادية اللغة. وقد اقتُرِحَت فكرة التَّرجمة الآلية الإحصائِيَّة في عام ١٩٤٩م، عندما فكَّرَ بعض العلماء في استخدام نظرية المعلومات وفك الشفرة لكتابة برامج الحواسيب لترجمة النص من لغة طبيعية إلى أخرى؛ وبعد أربعة عقود - في أواخر عام ١٩٨٠م، قامت مجموعة من باحثي شركة IBM بإعادة النظر في فكرة استخدام الأساليب الإحصائِيَّة للترجمة، وشجعهم على ذلك الزيادة في قوة الحوسبة، وتوافر كم كبير من النصوص المترجمة، وعدم إحراز تقدم ملحوظ في وسائل التَّرجمة الأخرى. وكانت طريقة التَّرجمة القائمة على أسس إحصائِيَّة هي النموذج الأبرز للترجمة الآلية في تسعينيَّات القرن العشرين والعقد الأول من القرن الحادي والعشرين لأسباب عديدة، منها: دقة التَّرجمة، وإمكانية تحسين التَّرجمة ببذل مجهود أقل من الطرق الأخرى، وكذلك سرعة بناء برنامج التَّرجمة للُغات متعددة متى توافر كمٌّ كبير من النصوص المترجمة للغتين.

في بدايات العقد الثاني من القرن الحادي والعشرين ظهرت تقنية استخدام التعلم العميق للترجمة الآلية؛ وقد أحرزت تقدمًا في استخدام الأساليب الإحصائِيَّة.

٢- تعريف بأهم المصطلحات المستَخدَمة في التَّرجمة الآلية

هناك بعض المصطلحات المستخدمة في كل تقنية من تقنيات التَّرجمة الآلية التي سبقت الإشارة إليها في القسم السابق (التقنية القائمة على قواعد لغوية، والتقنية القائمة على استخدام أمثلة لجمل سبق ترجمتها، والتقنية القائمة على أسس إحصائِيَّة باستخدام كم كبير من النصوص المتَرجمة).

٢,١- المصطلحات المستخدمة في التَّرجمة الآلية القائمة على القواعد اللغوية

منهج التَّرجمة المباشرة (Approach Direct):

وتعني ترجمة كل كلمة في لغة المصدر إلى ما يقابلها في لغة الهدف باستخدام قاموس ثنائي اللغة.

منهج التَّرجمة التحويلي (Approach Transfer):

ويعني تحليل جملة لغة المصدر ثم القيام بتحويل ناتج هذا التحليل إلى ما يقابله بلغة الهدف، وأخيرا توليد جملة لغة الهدف.

منهج التَّرجمة باستخدام اللغة الوسيطة (Approach Interlingua):

ويعني تحليل جملة لغة المصدر إلى لغة وسيطة تعتمد على مجال التَّرجمة وتعبر بطريقة منضبطة عن المعاني التي تحتويها جملة لغة المصدر مما يسهل توليد الجملة الممثلة باللغة الوسيطة إلى أي لغة أخرى.

قواعد اللغة الغير معتمدة على السياق (Grammar Free Context): وتتكوَّن من:

- مجموعة من الرموز النهائية؛ وتمثل مفردات اللغة (Terminals).
- مجموعة من الرموز الغير نهائية؛ وتمثل الوحدات البنيوية للغة (terminals-Non).
- مجموعة من القواعد التي تتكون من جانب أيمن وجانب أيسر. الجانب الأيمن يحتوي على رمز غير نهائي واحد، والجانب الأيسر يحتوي على مجموعة من الرموز الغير نهائية والرموز النهائية (Production Rules).
- رمز غير نهائي ابتدائي (Starting Symbol).

٢,٢- المصطلحات المستخدمة في التَّرجمة الآلية القائمة على استخدام أمثلة

مدونة ثنائية اللغة (Bilingual Corpus):

هي مجموعة كبيرة من النصوص بلغتين، إحدى هاتين اللغتين يطلق عليها لغة المصدر والأخرى يطلق عليها لغة الهدف. مجموعة النصوص بلغة الهدف هي ترجمة مجموعة النصوص بلغة المصدر دون أن تكون هناك محاذاة بين الجمل في مجموعتي النصوص.

مدونة ثنائية اللغة متوازية (Parallel Bilingual Corpus):

هي مدونة ثنائية اللغة، كل جملة بلغة الهدف تشير إلى جملة مكافئة لها بلغة المصدر.

برنامج التطابق (Matching Module):

هو البرنامج الذي يحاول العثور على أكبر عبارة في الجملة المدخلة تتطابق مع الأمثلة الموجودة في نصف المدونة الثنائية المتوازية المكتوبة باللغة التي يراد ترجمتها.

برنامج التعرف (Identification Module):

هو البرنامج الذي يحاول تحديد أفضل جزء يمكن اعتباره ترجمة للعبارة التي تم العثور عليها في الأمثلة الموجودة في المدونة الثنائية المتوازية في الجملة المقابلة لها.

برنامج تجميع العبارات (Assembling Module):

هو البرنامج الذي يحاول تجميع العبارات المترجمة لتكوين أفضل جملة.

٢,٣ - المصطلحات المستخدمة في التَّرجمة الآلية القائمة أسس إحصائِيَّة

مدونة أحادية اللغة (Mono Lingual Corpus):

مجموعة كبيرة من النصوص بلغة واحدة.

مُدَوَّنة مُتحاذية ثُنائِيَّة اللُّغة (Aligned Bilingual Corpus):

هي مدونة ثنائية اللغة، كل كلمة في جملة بلغة الهدف تشير إلى كلمة أو أكثر في الجملة المكافئة لها بلغة المصدر.

نموذج إحصائي للترجمة (Statistical Translation Model):

مجموعة من الاحتمالات المشروطة لترجمة كلمة أو عبارة من لغة المصدر إذا أعطيت كلمة أو عبارة من لغة الهدف.

نموذج إحصائي للغة (Language Model):

مجموعة من الاحتمالات المشروطة لظهور كلمة إذا ظهرت كلمة أو عدة كلمات سابقة لها.

قواعد السياق الحر المتزامن (Synchronous Context Free Grammar):

كل قاعدة من هذه القواعد تتكون من جانب أيمن يعبر عن مكون نحوي، وجانب أيسر يُمَثِّل مجموعة من الكلمات أو المكونات النحوية الأقل تعقيدا من الجانب الأيمن بلغة المصدر والمكافئ لها بلغة الهدف. ويتم توليد هذه القواعد من مدونة ثنائية متوازية ومتحاذية.

برنامج فاكّ الشفرة (Decoder):

برنامج يستخدم نموذجًا إحصائيًّا للترجمة ونموذجًا إحصائيًّا للغة الهدف ليولد لغة الهدف من لغة المصدر.

منهج التَّرجمة الإحصائي القائم على مستوى ترجمة الكلمة (Word based Statistical Machine Translation):

هو المنهج الذي يستخدم نموذج إحصائي للترجمة تكون الاحتمالات المشروطة فيه لترجمة كلمة من لغة المصدر إذا أعطيت كلمة من لغة الهدف.

منهج التَّرجمة الإحصائي القائم على مستوى ترجمة العبارة (Phrase Based Statistical Machine Translation):

هو المنهج الذي يستخدم نموذج إحصائي للترجمة تكون الاحتمالات المشروطة فيه لترجمة عبارة من لغة المصدر إذا أعطيت عبارة من لغة الهدف.

منهج التَّرجمة الإحصائي القائم على مستوى بناء هرمي للعبارة (Hierarchical Based Statistical Machine Translation):

هو المنهج الذي يستخدم نموذج إحصائي للترجمة مكون من مجموعة من قواعد السياق الحر المتزامن.

منهج التَّرجمة الإحصائي القائم على مستوى ترجمة لغة المصدر إلى بناء نحوي للغة الهدف (Syntax Based Statistical Machine Translation):

هو المنهج الذي يستخدم نموذجًا إحصائيًّا للترجمة مكوَّنًا من قواعد تربط بين الكلمات والعبارات والجمل من لغة المصدر مع الأشجار البنائية الناتجة عن التحليل اللغوي للجمل على جانب لغة الهدف.

٢, ٤ – المصطلحات المستخدمة في التَّرجمة الآلية القائمة على التعلم العميق

خلية عصبية (Neuron)

هي وحدة حاسوبية لها عدد من المدخلات ومخرج واحد، قيمته هي دالة في قيم مدخلاته.

شبكة عصبية (Neural Network)

هي مجموعة من الخلايا العصبية، مرتبة في طبقات، لها عدد من المدخلات وعدد آخر من المخرجات.

منهج التَّرجمة باستخدام الشبكة العصبية (Translation Machine Neural)

هو المنهج الذي يستخدم شبكة عصبية في الترجمة.

منهج التَّرجمة باستخدام التعلم العميق(Machine Translation using Deep Learning)

هو المنهج الذي يستخدم شبكة عصبية ذات طبقات متعددة، ويُستخدم أحيانًا مصطلح التَّرجمة باستخدام الشبكة العصبية للدلالة على نفس المنهج.

تمثيل الكلمة في مُتَّجه(Word Vector Rpresentation)

هي طريقة لتمثيل الكلمة في مُتَّجه رياضي عن طريق السِّياق الَّذي تظهر فيه الكلمة.

نموذج تسلسل الكلمات (Model Sequence Word)

في سياق الترجمة الآلية؛ هو تذكر تسلسل مجموعة من الكلمات في لغة مع ترجمتها إلى تسلسل من الكلمات في لغة أخرى.

شبكة عصبية متتالية(Recurrent Neural Network)

هي شبكة متتالية من الوحدات؛ تشتمِلُ كل وحدة بها على عدد محدد من الخلايا العصبية. ويكون لكل وحدة عدد من المدخلات وعدد من المخرجات. وتُضافُ مخرجات كل وحدة إلى مدخلات الوحدة التي تليها.

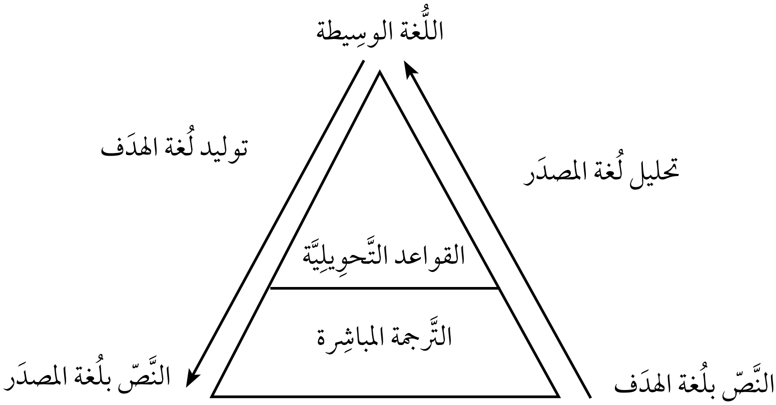
٣- تقنيات التَّرجمة الآلية، وآخر التَّوجُّهات البحثِيَّة

٣,١- التَّرجمة الآلية القائمة على القواعد اللغوية

الأقسام الفرعية التالية تصف ثلاثة مناهج لاستخدام القواعد اللغوية في التَّرجمة. المنهج الأول هو ما يطلق عليه منهج التَّرجمة المباشر، والمنهج الثاني هو ما يطلق عليه التَّرجمة باستخدام القواعد التحويلية، أما المنهج الثالث فهو ما يطلق عليه التَّرجمة من خلال تحويل جمل لغة المصدر إلى لغة وسيطة. والشكل (٢-١) يوضح الفروق بين المناهج الثَّلاثة من خلال رسم تخطيطي.

- **منهج التَّرجمة المباشر**

منهج التَّرجمة المباشر يقوم أساسا على استبدال كلمة مقابل كلمة بين زوج من اللغات باستخدام قاموس ثنائي اللغة واستخراج مقابلها من اللغة المستهدفة. وعيوب هذا النظام عدم قدرته على تحليل البنية النحوية أو العلاقات الدلالات في جمل الإدخال مما يسفر عن قلة جودة التَّرجمة. كما أن بناء القاموس ثنائي اللغة يجب أن يحتوي على كم كبير من الكلمات بكل تصريفاتها باللغتين، وذلك لكل زوج من اللغات؛ وبناء هذه القواميس مكلف للغاية.



الشَّكل ٢-١: مناهج التَّرجمة القائمة على القواعد اللغوية.

- **منهج التَّرجمة باستخدام القواعد التحويلية**

منهج التَّرجمة باستخدام القواعد التحويلية يمثل حلاًّ وسطا بين منهج التَّرجمة المباشر ومنهج التَّرجمة من خلال تحويل جمل لغة المصدر إلى لغة وسيطة. هذا المنهج يعمل على ثلاث مراحل: تحليل جمل لغة المصدر لغويا: صرفيا أو نحويا أو دلاليا، ثم تحويل ناتج التحليل إلى مقابل له في لغة الهدف، وبعد ذلك يتم توليد جمل لغة الهدف المكافئة باستخدام قواعد الصرف والنحو للغة المستهدفة. هذا المنهج يتكون من ثلاثة مكونات. المكون الأول خاص بلغة المصدر ويحتوي على قاموس للغة المصدر ومحللات صرفية ونحوية ودلالية لها، والمكون الثاني يحتوي على قاموس ثنائي اللغة وقواعد تحويلية تربط بين الوحدات المعجمية والنحوية لجمل لغة المصدر والوحدات المقابلة لها من لغة الهدف، أمَّا المكون الثالث فيحتوي على قاموس للغة الهدف ومولدات صرفية ونحوية ودلالية لها. عيب هذا المنهج أنه يجب إعادة كتابة المكون الثاني لكل زوج من اللغات مما يحتاج إلى تكلفة كبيرة؛ وهناك عيب آخر يتمثَّل في أن معالجة الالتباس اللغوي الذي يتم على لغة المصدر لا تُحَقِّق النَّتائج المنشودة - حتى في أفضل وضع وهو القيام بكل التحليلات الممكنة وصولا إلى التحليل الدلالي، حيث إن كثيرًا من أنواع الالتباس لا يمكن حلها إلا من خلال السياق الذي ذكرت فيه الجمل المطلوب ترجمتها، وكذلك المعرفة عن العالم التي يصعب تمثيلها ومعالجتها آليا.

- **منهج التَّرجمة القائم على اللغة الوسيطة**

منهج التَّرجمة القائم على اللغة الوسيطة (إنترلينجو). ويتألف نظام إنترلينجو من مكونين: المكون الأول هو الذي يحلل النص المكتوب بلغة المصدر ويحوله إلى تمثيل مقابل بلغة مستقلة مجردة، وهي ما نطلق عليها اللغة الوسيطة، والمكون الآخَر هو الذي يولد النص المكافئ للنص الأصلي بلغة الهدف من اللغة الوسيطة. في هذا المنهج لا يوجد اتصال بين المكون الأول الذي يقوم بتحليل النص الأصلي والمكون الآخَر الذي يقوم بتوليد النص الأصلي بلغة أخرى؛ وعلى الرغم من مزايا هذا المنهج الذي يقدم حلولا لأغلب المشكلات التي يعاني منها المنهج القائم على القواعد التحويلية إلا أنه لم يُستخدم على نطاق

واسع لسبَبين، أحدهما: صعوبة تعريف لغة محايدة وسيطة بين لغات متباينة، والآخرُ: صعوبة أن تكون هذه اللغة خالية من أي التباس وقادرة على تمثيل أي محتوى لنص مكتوب بلغة طبيعية.

مما تقدم نستطيعُ أن نقولَ إن منهج التَّرجمة القائم على القواعد التحويلية هو المنهج الأكثر استخداما بين المناهج القائمة على القواعد اللغوية، حيث إن هذا المنهج يقدم حلا وسطا بين منهج التَّرجمة المباشر، والذي يعاني من سوء جودة التَّرجمة مع بساطته التقنية، ومنهج التَّرجمة القائم على اللغة الوسيطة، والذي يتميز بالقدرة على التعامل مع أزواج كثيرة من اللغات مع صعوبة تعريف لغة وسيطة تستوعب كل الصور التي يمكن التعبير عنها باللغات الطبيعية.

٣,٢- طريقة التَّرجمة الآلية القائمة على استخدام أمثلة

تتميز الطريقة القائمة على استخدام الأمثلة للترجمة الآلية بأنها تستخدم مدونة ثنائية اللغة متوازية، وتمثل هذه المدونة قاعدة معرفية لبرنامج التَّرجمة. الفكرة الأساسية لهذه الطريقة هي التَّرجمة من خلال التماثل في التكوين الظاهري للجملة، وليس من خلال القيام بتحليل لغوي عميق لها؛ ومرجع هذه الفكرة هو الاعتقاد بأن الناس تقوم أولا بتحليل الجملة إلى عبارات ثم تقوم بترجمة هذه العبارات، وأخيرا تُكَوَّن الجملة بشكل صحيح من العبارات المترجمة.

وتترجم العبارات عن طريق التطابق مع عبارات سبق ترجمتها موجودة في مجموعة النصوص المترجمة المتوازية. ويتكون نظام التَّرجمة القائم على استخدام أمثلة من الأجزاء التالية:

- برنامج التطابق الذي يحاول العثور على أكبر عبارة في الجملة المدخلة تتطابق مع الأمثلة الموجودة في نصف المدونة ثنائية اللغة المتوازية المكتوبة بنفس لغة الجملة المدخلة، أي التي يُراد ترجمتها.
- برنامج التعرف الذي يحاول تحديد أفضل جزء يمكن اعتباره ترجمة للعبارة التي تم العثور عليها في الجملة الموجودة في النصوص المترجمة المتوازية في الجملة المقابلة لها.
- برنامج تجميع العبارات المترجمة لتكوين أفضل جملة.

٣,٣- طريقة التَّرجمة الآلية القائمة على أسس إحصائِيَّة

هذه الطريقة الإحصائِيَّة تقوم على بناء نموذج إحصائي للترجمة ونموذج إحصائي للغة، ليستخدما بعد بنائهما بواسطة برنامج لتوليد لغة الهدف من لغة المصدر؛ وهذا البرنامج يطلق عليه «برنامج فاكّ الشفرة». وهذا الاسم قد تم إطلاقه على هذا البرنامج لأسباب تاريخية، إذ إنه في بداية الأبحاث في التَّرجمة الآلية كان يُنظر إليها على أن جملة لغة الهدف تم تشفيرها إلى لغة المصدر وأن المترجم الآلي هو الذي يقوم بفك جملة المصدر المشفرة إلى جملة الهدف. والاختلاف بين مناهج التَّرجمة على أسس إحصائِيَّة قائم على طريقة بناء نموذج التَّرجمة الإحصائي، ومن ثم على كتابة البرنامج المناسب لاستخدام هذا النموذج لبرنامج فاكّ الشفرة.

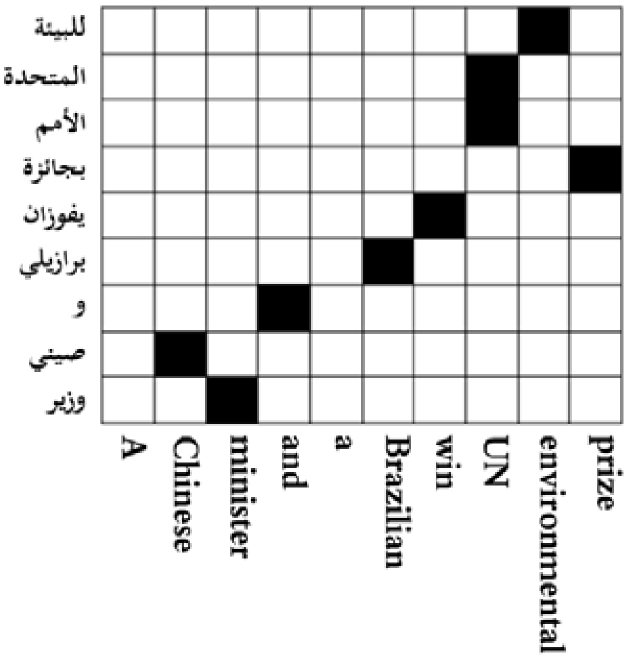
وسنعرِضُ في هذا القسم لمناهج التَّرجمة القائمة على أسس إحصائِيَّة، والَّتي تتمثَّلُ في: المنهج القائم على مستوى ترجمة الكلمة، والمنهج القائم على مستوى ترجمة العبارة، والمنهج القائم على مستوى ترجمة بناء هرمي للعبارة، والمنهج القائم على مستوى ترجمة لغة المصدر إلى بناء نحوي للغة الهدف.

- **المنهج القائم على مستوى ترجمة الكلمة**

في النماذج القائمة على ترجمة كلمة، يكون نموذج التَّرجمة عبارة عن مجموعة من الاحتمالات لترجمة كلمات من لغة المصدر إلى كلمات من لغة الهدف، ويتم تقدير هذه الاحتمالات من مدونة ثنائية متحاذية. هناك خمسة نماذج أساسية لتقدير ترجمة كل كلمة من لغة الهدف إلى أكثر من كلمة في لغة المصدر؛ وهذه النماذج الخمسة تم اقتراحها من قِبَل مركز أبحاث IBM في بداية التسعينيَّات من القرن الماضي. تعتمد هذه النماذج على فرض أن كل كلمة في جملة في لغة الهدف قد يكون مصدرها أي كلمة في جملة لغة المصدر الموازية لها، حيث يتم توليد جميع التباديل للكلمات في كل جملتين في المدونة المتحاذية، وكل تبديل من هذه التباديل يعطى احتمالاً متساويًا في البداية.

وباستخدام هذه التباديل يتم حساب احتمالات ترجمة كل كلمة من كلمات لغة الهدف إلى ما يقابلها من كلمات في لغة المصدر؛ وبناء على هذه الاحتمالات يُعادُ حساب احتمالات التباديل المختلفة لكل جملة حتى يتم الوصول إلى أفضل تقابل بين كل كلمة

في جملة لغة الهدف والكلمة أو الكلمات المقابلة لها في الجملة المحاذية من لغة المصدر لكل الجمل في المدونة. الشكل (٢-٢) يوضح مثالاً لجملتين متوازيتين بعد تحديد التقابل بين الكلمات في الجملتين.



الشَّكل ٢-٢: مثال لجملة تم تحديد تقابل الكلمات بها.

(يحتوي ملحق هذا الباب على تفصيل أكثر حول طريقة بناء هذا النموذج الإحصائي).

والمشكلة الأساسيَّة في النَّماذج القائمة على ترجمة كلمة أنها لا تستخدم أي كلمات سابقة أو لاحقة لتقدير احتمال التَّرجمة، وفي بعض الأحيان يكون من المفيد أن تكون وحدة التَّرجمة أكثر من كلمة لإنتاج ترجمة جيدة. فعلى سبيل المثال، إذا أردنا ترجمة «وزير صيني» إلى اللغة الإنجليزية – مع افتراض أنه قد تم تحديد أن كلمة «وزير» معناها (Minister) وأن كلمة «صيني» معناها (Chinese)، فإن التَّرجمة سوف تكون (Minster Chinese)، وهذه ترجمة خاطئة؛ أمَّا إذا كانت وحدة التَّرجمة مكونة من كلمتين وتم تحديد ترجمة هذه الوحدة، فإننا سوف نحصل على التَّرجمة الصحيحة.

- **المنهج القائم على مستوى ترجمة العبارة**

تم اقتراح هذا المنهج للتغلب على المشكلات الناتجة عن المنهج القائم على مستوى ترجمة الكلمة. وحدة التَّرجمة في هذا المنهج هي مجموعة من الكلمات المتلاصقة. هذه المجموعة من الكلمات المتلاصقة – والتي سوف نطلق عليها عبارة لا تمثل أيَّ مكون لغوي – ليست سوى سلاسل من الكلمات المختارة وفقا لمحاذاة كل كلمة في جملة المصدر لمقابلها في جملة الهدف. فعلى سبيل المثال، يمكن توليد العبارات المكونة من كلمتين وترجمتها من المصفوفة الموضحة في الشكل (٢–٢) كما في الجدول (٢–١):

العبارة العربية	التَّرجمة الإنجليزية
وزير صيني	Chinese minister
برازيلي يفوزان	Brazilian win
الأمم المتحدة	UN

الجدول ٢–١: العبارات المتقابلة التي يمكن توليدها من المصفوفة الموضحة في الشَّكل رقم (٢–٢).

ومن خلال المدونة اللغوية المتوازية – والتي تم مقابلة كل كلمة في جملها المكتوبة بلغة الهدف إلى الكلمة المقابلة لها بلغة المصدر – يتم حساب احتمالات ترجمة كل العبارات بأطوالها المختلفة: الأحادية، الثنائية، الثلاثية، ... إلخ من لغة الهدف إلى لغة المصدر.

- **المنهج القائم على مستوى ترجمة بناء هرمي للعبارة**

يقوم هذا المنهج على استخدام قواعد السياق الحر المتزامن؛ وكل قاعدة من هذه القواعد تتكون من جانب أيمن يعبر عن مكون نحوي، وجانب أيسر يُمَثِّل مجموعة من الكلمات أو المكونات النَّحوية الأقل تعقيدا من الجانب الأيمن بلغة المصدر والمكافئ لها بلغة الهدف. فعلى سبيل المثال يمكن تمثيل العبارات المتقابلة الموضحة في الجدول (٢–١) في قواعد سياق حر متزامن كما يلي:

(١) X → <وزير صيني, Chinese minister>

(٢) X → < برازيلي يفوزان, Brazilian win>

(٣) X → < الأمم المتحدة, UN>

من هذه القواعد، ومن الجملتين السابق استخدامهما في الشكل (١١-٢)، يمكن توليد القاعدة التالية:

(٤) $X \rightarrow$ < X_1 و X_2 بجائزة X_3 للبيئة, A X_1 and a X_2 X_3 environmental prize >

ولاستكمال هذا المثال، يجب إضافة القاعدتين التاليتين حتى يمكن توضيح كيفية عمل برنامج فاكّ الشفرة لهذا النموذج:

(٥) $S \rightarrow$ < S_1 X_2, S_1 X_2 >

(٦) $S \rightarrow$ < X_1, X_1 >

تتمّ عملية فك الشفرة من خلال البدء بالرمز الابتدائي للقواعد الحرة المتزامنة والجملة المراد ترجمتها ثم محاولة تطبيق القواعد الأخرى للحصول على ترجمة الجملة المطلوبة. وفيما يلي خطوات تطبيق القواعد لترجمة الجملة العربية في المثال المستخدم لتوضيح الفكرة:

S > $\rightarrow$ X_1 , X_1 >

$\rightarrow$ < X_1 و X_2 بجائزة X_3 ، للبيئة A X_1 and a X_2 X_3 environmental prize >

$\rightarrow$ < وزير صيني و X_2 بجائزة X_3 للبيئة، A Chinese minister and a X_2 X_3 environmental prize >

$\rightarrow$ < وزير صينيو برازيلي يفوزان بجائزة X_3 للبيئة، A Chinese minister and a Brazilian win X_3 environmental prize >

$\rightarrow$ < وزير صينيو برازيلي يفوزان بجائزة المتحدة الأمم للبيئة ، A Chinese minister and a Brazilian win UN environmental prize >

باختصار فإنَّ البرنامج المقترح لفك الشفرة هو برنامج بحث ذكي لاختيار أفضل القواعد التي يجب تطبيقها لترجمة جملة بلغة المصدر إلى جملة بلغة الهدف، حيث إنه في الواقع يكون هناك أكثر من ترجمة لجملة بلغة المصدر.

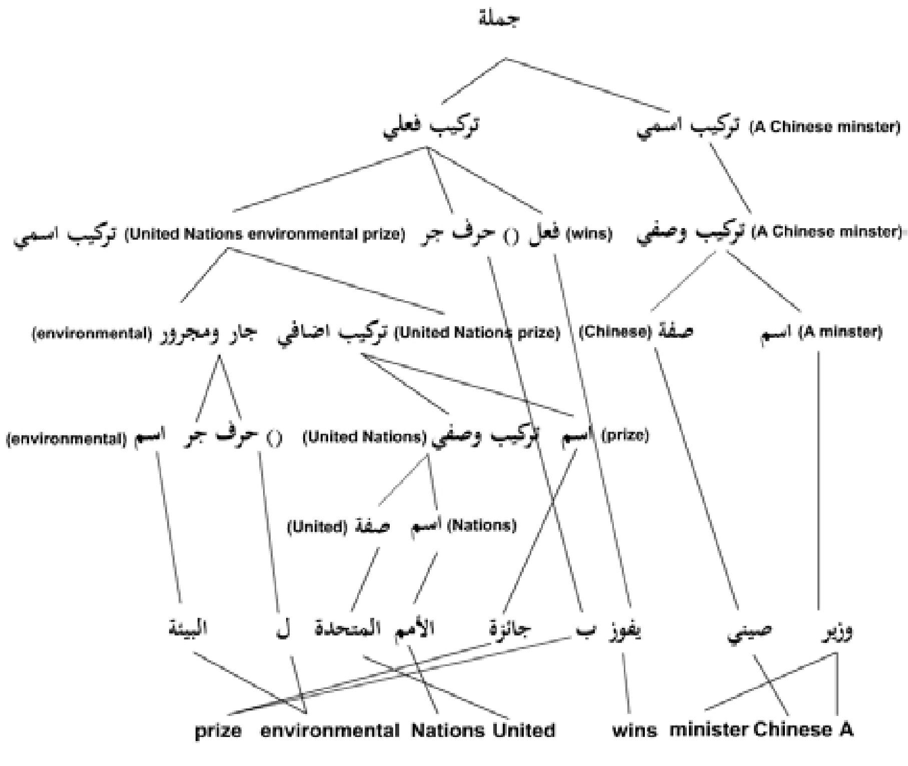
- **المنهج القائم على مستوى ترجمة لغة المصدر إلى بناء نحوي للغة الهدف**

تقوم فكرة هذا المنهج على التحليل النَّحوي للجمل على جانب لغة الهدف في المدونة ثنائية اللغة، ومحاذاة الكلمات من كلا الجانبين، ثم تعلم قواعد ترجمة تربط بين الكلمات والعبارات والجمل من لغة المصدر مع الأشجار البنائية الناتجة عن التحليل اللغوي للجمل على جانب لغة الهدف. هذه المجموعة من قواعد التَّرجمة تعتبر نموذج التَّرجمة للمنهج القائم على مستوى ترجمة لغة المصدر إلى بناء نحوي للغة الهدف؛ ويستند برنامج فاكّ الشفرة في هذا المنهج إلى بناء شجرة

التحليل البنيوي لجملة الهدف، لجملة مدخلة بلغة المصدر، باستخدام نموذج التَّرجمة الذي تم بناؤه، ثم تحويلها إلى البناء الظاهري للغة الهدف. لتوضيح هذه الفكرة سوف نعطي مثالا لبناء نموذج التَّرجمة من الإنجليزية إلى العربية. الشكل (٢-٢) يصف مخطط محاذاة للجملة (مع ملاحظة أننا قمنا بكتابة اللغة الإنجليزية من اليمين إلى اليسار حتى يمكن رسم خطوط المحاذاة بصورة أفضل مما يمكن القارئ من متابعتها):

(A Chinese minister wins United Nations environmental prize)

وترجمتها إلى اللغة العربية: «وزير صيني يفوز بجائزة الأمم المتحدة للبيئة»



الشكل ٢-٣: مخطط محاذاة لجملة إنجليزية والشجرة البنيوية لترجمتها العربية.

من هذا المخطط الموضح في الشكل رقم (٢-٣) يمكن استخراج قواعد التَّرجمة لعبارات باللغة الإنجليزية إلى شجرة بنيوية باللغة العربية كما في الشكل رقم (٢-٤).



الشَّكل ٢-٤: بعض القوعد المستخلصة من المخطط الموضح في الشَّكل (٢-٣).

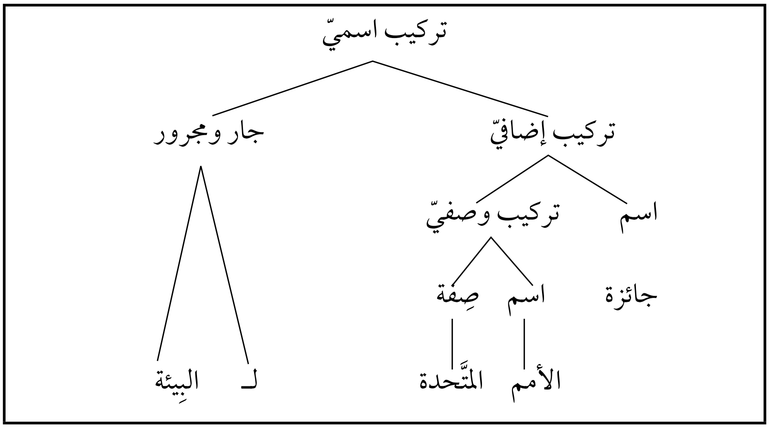
ويُمكن تقسيم القواعد المستخلصة في الشكل (٢-٤) إلى ثلاث أنواع:

- قواعد خاصة، مثل القاعدتين (أ) و (ب)، حيث إن مدخلات كل منهما كلمات باللغة الإنجليزية ومخرجات كل منهما شجرة بنيوية باللغة العربية. أوراق هذه الشجرة ترجمة العبارة المدخلة باللغة العربية مرتبة ترتيبا نحويا صحيحا.
- قواعد عامة، مثل القاعدتين (هـ) و (و)، حيث إن مدخلات كل منهما رمز أو رموز غير نهائية، ومخرجات كل منهما شجرة بنيوية باللغة العربية تربط المدخلات في شجرة واحدة.

- قواعد شبه عامة مثل القاعدتين (ج) و (د)، حيث إن مدخلات كل منهما رمز أو رموز غير نهائية مع رموز نهائية، ومخرجات كل منهما شجرة بنيوية باللغة العربية تربط المدخلات في شجرة واحدة.

ولتوضيح فكرة عمل برنامج فاكّ الشفرة لهذا المنهج، سنُحاول أن نتتبَّع خطوات ترجمة المثال المستخدم لتوليد هذه القواعد على النَّحو التَّالي:

- يتم قراءة الجملة المدخلة كلمة كلمة حتى يمكن اختيار قاعدة تكون مدخلاتها الكلمات المقروءة. الكلمات المقروءة في هذا المثال – والتي تتطابق مع مدخلات القاعدة (أ) الموضحة في الشكل (٢–٤) – هي: «Minister Chinese A» والتي تولد التركيب الوصفي المكون من الكلمتين «وزير صيني».
- يتم تكرار الخطوة الأولى حتى يتم الانتهاء من تطبيق القواعد الخاصة. في حالة هذا المثال.. فإن القاعدة (ب) هي التي سوف تطبق ويتم توليد تركيب وصفي آخر مكون من الكلمتين «الأمم المتحدة».
- القاعدة (د) يمكن تطبيقها الآن والتي تغطي التركيب الوصفي (United Nations)، والتي يتبعها (Prize Environmental)؛ وسوف ينتج عن هذا التطبيق توليد الشجرة الموجودة في هذه القاعدة بعد تركيب الشجرة الفرعية للتركيب الوصفي، كما هو موضح في الشكل رقم (٢–٥).



الشَّكل ٢–٥: الشجرة البنيوية للعبارة «جائزة الأمم المتحدة».

- القاعدة (ج) يمكن تطبيقها بعد تكوين التركيب الاسمي في الخطوة رقم (3)، والتي تولد الشجرة البنيوية للعبارة «يفوز بجائزة الأمم المتحدة للبيئة»، وهي تمثل تركيبًا فعليًّا.
- القاعدة (هـ) هي الوحيدة التي يمكن تطبيقها في هذا الموقف، والتي تحوِّل التركيب الوصفي «وزير صيني» إلى تركيب اسميّ.
- تم الآن تكوين تركيب اسمي يتبعه تركيب فعلي، ومن ثم يمكن تكوين الشجرة البنيوية للجملة باستخدام القاعدة (و). ويمكن لبرنامج فاكّ الشفرة أن يولد جملة الهدف باللغة العربية مرتبة ترتيبًا نحويًّا صحيحًا.

٣,٤- طريقة التَّرجمة الآلية القائمة على التعلم العميق

طريقة الترجمة الآلية القائمة على التعلم العميق أبسط من طرق الترجمة القائمة على النموذج الإحصائي؛ ذلكَ أَنَّهُ لا يوجد نموذج لغوي منفصل، ولا نموذج للترجمة، ولا نموذج فاك للشفرة. هذه الطريقة تعتمد على تعليم شبكة عصبية متتالية بوضع جمل لغة المصدر كمدخلات لهذه الشبكة ووضع ترجمة هذه الجمل بلغة الهدف كمخرجات لها. وتحتاجُ عملية التعلم هذه إلى قوة حاسوبية عالية. وحالَ القُدرة على تعليم هذه الشبكة العصبية، يتم استخدامها في ترجمة أي جمل أخرى من لغة المصدر.

٤- البرامج والموارِد اللغوية المرتبطة بالتَّرجمة الآلية

٤,١- البرامج اللغوية المطلوبة لنظم التَّرجمة الآلية القائمة على القواعد اللغوية

تحتاج مناهج التَّرجمة القائمة على القواعد اللغوية إلى برامج على جانب لغة المصدر للقيام بالتحليل الصرفي والتحليل النحوي والتحليل الدلالي وتوليد اللغة الوسيطة؛ كما تحتاج على جانب لغة الهدف إلى برامج للقيام بتحويل الجملة الممثلة باللغة الوسيطة إلى تمثيل دلالي، وتحويل التمثيل الدلالي إلى تمثيل نحوي، وتحويل التمثيل النحوي إلى جذوع مزيدة بخصائصها الصرفية، وتركيب هذه الجذوع مع خصائصها الصرفية لتكوين الكلمات في صورتها النهائية. وفي حالة منهج التَّرجمة باستخدام القواعد التحويلية فسنكون في حاجة إلى قواعد تحويلية تتناسب مع مستوى التحليل الذي يتم على جانب

لغة المصدر. وفي أغلب الأحيان يتم تحليل النص المكتوب بلغة المصدر نحويا ويتم تمثيل ناتج التحليل في شجرة بنائية. وفي هذه الحالة يتم كتابة برنامج لتحويل هذه الشجرة البنائية التي تمثل الجملة المراد ترجمتها إلى شجرة مكافئة بلغة الهدف. والجدول (٢-٢) يلخص العلاقة بين مناهج التَّرجمة القائمة على القواعد اللغوية والبرامج المطلوبة.

	محلل صرفي	محلل نحوي	محلل دلالي	مولد لغة وسيطة	مولد دلالي	مولد نحوي	مولد صرفي	مركب كلمات	مكون تحويلي
منهج الترجمة المباشر									X
منهج الترجمة باستخدام القواعد التحويلية	X	X					X	X	X
منهج الترجمة القائم على اللغة الوسيطة	X	X	X	X	X	X	X	X	X

الجدول ٢-٢: العلاقة بين مناهج التَّرجمة القائمة على القواعد اللغوية والبرامج المطلوبة لتنفيذ هذه المناهج

٤ , ٢- البرامج اللغوية المطلوبة لنظم التَّرجمة الآلية القائمة على الأمثلة

البرامج الأساسية التَّرجمة الآلية القائمة على الأمثلة هي:

- برنامج التطابق الذي يحاول العثور على أكبر عبارة في الجملة المدخلة تتطابق مع الأمثلة الموجودة في نصف المدونة الثنائية المتوازية المكتوبة باللغة التي يراد ترجمتها.
- برنامج التعرف الذي يحاول تحديد أفضل جزء يمكن اعتباره ترجمة للعبارة التي تم العثور عليها في الجملة الموجودة في النصوص المترجمة المتوازية في الجملة المقابلة لها.
- برنامج تجميع العبارات المترجمة لتكوين أفضل جملة.

ويحتاج هذا المنهج إلى الموارِد التالية:

- مدونة ثنائية اللغة متوازية، وقاموس ثنائي اللغة لبرنامج التطابق، وبرنامج التعرف.
- نموذج اللغة المستهدفة لبرنامج تجميع العبارات.

٤, ٣- البرامج اللغوية المطلوبة لنظم التَّرجمة الآلية القائمة على أسس إحصائِيَّة

البرامج الأساسية للترجمة الآلية القائمة على أسس إحصائِيَّة هي:

- برنامج محاذاة الكلمات.
- برنامج بناء نموذج ترجمة العبارة.
- برنامج توليد قواعد لغوية متحررة من السياق ومتزامنة.
- برنامج توليد قواعد ترجمة.
- برنامج فاكّ الشفرة الخاص بالنظام القائم على مستوى ترجمة الكلمة.
- برنامج فاكّ الشفرة الخاص بالنظام القائم على مستوى ترجمة العبارة.
- برنامج فاكّ الشفرة الخاص بالنظام القائم على مستوى بناء هرمي للعبارة.
- برنامج فاكّ الشفرة الخاص بالنظام القائم على مستوى ترجمة لغة المصدر إلى بناء نحوي للغة الهدف.

الجدول (٢-٣) يوضح العلاقة بين مناهج التَّرجمة القائمة على أسس إحصائِيَّة والبرامج المطلوبة لتنفيذ هذه المناهج.

برنامج بناء نموذج اللغة

	برنامج محاذاة الكلمات	برنامج بناء نموذج ترجمة العبارة	برنامج توليد قواعد لغوية متحررة من السياق ومتزامنة	برنامج توليد قواعد ترجمة	برنامج فاكّ الشّفرة	برنامج بناء نموذج اللغة
المنهج القائم على مستوى ترجمة الكلمة	X				برنامج فاكّ الشَّفرة الخاصّ بالنِّظام القائم على مستوى ترجمة الكلمة	X
المنهج القائم على مستوى ترجمة العبارة	X	X			برنامج فاكّ الشَّفرة الخاصّ بالنِّظام القائم على مستوى ترجمة العبارة	X
المنهج القائم على مستوى ترجمة بناء هرمي للعبارة	X		X		برنامج فاكّ الشَّفرة الخاصّ بالنِّظام القائم على مستوى ترجمة بناء هرمي للعبارة	X
المنهج القائم على مستوى ترجمة لغة المصدر إلى بناء نحوي للغة الهدف	X			X	برنامج فاكّ الشَّفرة الخاصّ بالنِّظام القائم على مستوى ترجمة لغة المصدر إلى بناء نحوي للغة الهدف	X

الجدول ٢-٣: العلاقة بين مناهج التَّرجمة القائمة على أسس إحصائِيَّة والبرامج المطلوبة لتنفيذ هذه المناهج

٥- أهم المواقع والأدوات المساعدة للموارِد والتِّقنيات مفتوحة المصدر

معظم البرامج والموارِد اللغوية لنظم التَّرجمة الآلية القائمة على القواعد اللغوية ليست متاحة للجمهور؛ إلا أن هناك بعض الشركات العاملة في مجال تقنيات اللغة العربية تمتلك محللات صرفية ومعاجمَ للغة العربية. كما تضمنت بعض الأطروحات في الجامعات قواعد نحوية مزيدة بملامح دلالية وقواعد لتوليد اللغة من لغة وسيطة؛

والمراجع في نهاية الكتاب تحتوي على هذه الأطروحات . ولكن هذه البرامج والموارِد التي تم تطويرها ليست متاحة للجمهور أيضا. أمَّا بالنسبة للترجمة الآلية القائمة على أسس إحصائِيَّة، فلا توجد الكثير من الموارِد المتاحة للجمهور بدون مقابل. إلا أن مؤسَّسة LDC في جامعة بنسلفانيا بالولايات المتحدة الأمريكية تتيح بعض موارِدها بدون مقابل للمتنافسين في المسابقة التي يجريها المعهد القومي للمعايرة والتقنية بالولايات المتحدة الأمريكية. وفيما يلي قائمة بالموارِد التي تمت إتاحتها للمتنافسين في المسابقة التي أجراها المعهد القومي للمعايرة والتقنية عام ٢٠٠٩ على التَّرجمة الآلية من العربية إلى الإنجليزية:

1- LDC2007T40 Arabic Gigaword Third Edition.

2- LDC2004T18 Arabic English Parallel News Part 1.

3- LDC2004T17 Arabic News Translation Text Part 1.

4- LDC2005E46 Arabic Treebank English Transla-tion.

5- LDC2005T02 Arabic Treebank: Part 1 v 3.0 (POS with full vocalization + syntactic analysis).

6- LDC2004T02 Arabic Treebank: Part 2 v 2.0.

7- LDC2005T20 Arabic Treebank: Part 3 (full corpus) v2.0 (MPG + Syntactic Analysis).

8- LDC2004L02 Buckwalter Arabic Morphological Analyzer.

.9- LDC2007T07 English Gigaword Third Edition.

10- LDC2004E72 eTIRR Arabic English News Text.

11- LDC2003T18 Multiple-Translation Arabic (MTA) Part 1.

12- LDC2005T05 Multiple-Translation Arabic (MTA) Part 2.

13- LDC2006E44 TIDES MT 2004 Arabic evaluation data.

14- LDC2006E39 TIDES MT 2005 Arabic evaluation data.

15- LDC2004E13 UN Arabic English Parallel Text.

أما البرامج والأدوات التي تُستخدم على نطاق واسع من الباحثين المهتمين بالتَّرجمة الآلية القائمة على أسس إحصائِيَّة والمتاحة للجمهور فهي:

٥,١- **جيزة ++ (GIZA++)**. هو امتداد للبرنامج الجيزة الذي تم تطويره خلال صيف عام ١٩٩٩ أثناء ورشة عمل في مركز اللغات في جامعة جونز هوبكنز. الجيزة ++ يستخدم من قبل العديد من العلماء لبناء نموذج التَّرجمة القائم على مستوى الكلمة، كما يُستخدم لمحاذاة الكلمات في مدونة ثنائية اللغة. ويمكن تحميل هذه الأداة مجانا من شبكة الإنترنت(١).

٥,٢- **البرمجيات المتاحة مجانا** من جامعة كارنيجي ميلون-كامبريدج لبناء نماذج إحصائِيَّة للغات(٢). وكذلك البرمجيَّات المتاحة من جامعة ستانفورد لنفس الغرض(٣).

٥,٣- **هناك مجموعة من برامج فك التشفير المتاحة مجانا**، والتي يمكن تحميلها. فهناك برنامج فاكّ الشفرة للنظام المبني للترجمة على مستوى العبارة ويسمى (Pharaoh)(٤)، كما أن هناك برنامجًا يستخدم بكثرة هذه الأيام يسمى (Moses)(٥).

٦- أفكارٌ لتطوير مدونات لغوية مستقبلية

حيث إنَّ أكثرَ المدونات ثنائية اللغة غير متاحة مجانا للباحثين، كما أن المتاح منها بمقابل في مجال الأخبار فقط، فإن هناك احتياج لاستحداث مدونات ثنائية اللغة في مجالات أخرى. ويفضل اختيار المجالات التي بها مادة مترجمة إلى أكثر من لغة، مثل: مواقع الأمم المتحدة على شبكة الويب. كذلك يمكن إضافة قيمة للمدونات ثنائية اللغة المتاحة لتحسين جودة نموذج الترجمة الناتج عن هذه المدونات.

1- http//:www.fjoch.com/GIZA.++html.

2- http://mi.eng.cam.ac.uk/~prc14/toolkit.html.

3- http://www.speech.sri.com/projects/srilm/.

4- http://www.isi.edu/licensed-sw/pharaoh/.

5- http://sourceforge.net/projects/mosesdecoder/.

٦ , ١- موضُوع الفِكرة الأولى:

تذييل مُدَوَّنة ثنائية اللغة صرفيًّا ودلاليًّا

- مادة الدراسة:

مدونة ثنائيَّة اللُّغة.

- الأسئِلة البَحثِيَّة:

– ما هي مجموعة المعلمات/ الرُّموز (tags) التي تُستخدم لتذييل الكلمات؟

– ما هي المنهجية المناسبة لتذييل الكلمات في المدونة ثنائية اللغة ؟

- منهج الدِّراسة، ومجال البحث:

تقوم الدراسة على استخدام برمجيات لمساعدة الباحث في تذييل الكلمات في الجمل المتقابلة، والتي قد تصل إلى خمسين ألف جملة (حوالي مليون كلمة) على الأقل؛ ومن ثم يمكن تحسين نموذج الترجمة الإحصائي الذي يمكن إنتاجه من هذه المدونة.

٦ , ٢- موضُوع الفِكرة الثَّانية:

بناء مدونة متعددة اللغات في مجالات منظمات الأمم المتحدة

- مادة الدراسة:

مواقع منظمات الأمم المتحدة الإلكترونيَّة، والتي تحتوي على وثائق متعددة اللغات؛ أو استخدام بعض الكتب المترجمة المتاحة.

- الأسئِلة البَحثِيَّة:

– ما هي المنهجية المناسبة لمحاذاة الجُمل في المدونة متعددة اللغات، حيث إنَّ الترجمات قد لا تكون حرفية؟

– ما هو الأسلوب الأمثل لتعظيم الفائدة من بناء هذه المدونة متعددة اللغات، حيث إنَّ حجمها قد لا يكون كبيرًا؟

- منهج الدِّراسة، ومجال البحث:

تقوم الدراسة على استخدام برمجيات لمساعدة الباحث لمحاذاة الجمل المتقابلة، والتي قد تصل إلى خمسين ألف جملة (حوالي مليون كلمة) على الأقل مع محاذاة ١٠٪ من هذه الجمل على مستوى الكلمة، وذلك لتحسين عملية محاذاة الكلمات تلقائيا بواسطة برمجيات مثل برنامج Giza ++، ومن ثم يمكن تحسين نموذج الترجمة الإحصائي الذي يمكن إنتاجه من هذه المدونة.

ملحق - الأساس النَّظريّ لبناء نظام ترجمة آليّ إحصائيّ

يوضح هذا الملحق الأساس النظري لبناء نظام ترجمة آلي إحصائي؛ ويرجع هذا الأساسُ إلى نظرية «القناة المشوشة» المعروفة في حقل المعلومات. يقوم تطبيق هذه النظرية في الترجمة الآلية على تصور أن الجملة الأصلية قد تم إرسالها من مصدر في قناة اتصال ووصلت مشوشة إلى هدفها؛ هذه الجملة المشوشة هي ترجمة الجملة الأصلية. وعملية الترجمة هي إرجاع الجملة المشوشة إلى أصلها. يمكن التعبير عن عملية الترجمة باستعمال نظرية الاحتمالات كما يلي:

$$\underset{e}{\operatorname{argmax}}\ P(e \mid f) \quad (١)$$

إذا افترضنا أن الحرف «e» يشير إلى جملة باللغة العربية وأن الحرف «f» يشير إلى أي لغة أجنبية وأن هناك أكثر من ترجمة للجملة «f» وأن لكل ترجمة قيمة مختلفة للتعبير الاحتمالي $P(e \mid f)$، فإنه يمكن قراءة التعبير الاحتمالي المذكور أعلاه كما يلي: الجملة العربية «e» التي تنتج أكبر قيمة للتعبير الاحتمالي $P(e \mid f)$ تكون هي الترجمة الأكثر احتمالا للجملة «f». وإذا افترضنا أن:

١-عدد الكلمات في الجملة هو «m».

٢-الجملة «f» مكونة من الكلمات «$f_1 \ldots\ldots f_j \ldots\ldots\ldots\ldots\ldots\ldots f_{m-1}.f_m$».

٣-كل كلمة «f_j» يمكن أن تترجم إلى أكثر من ترجمة. ولنقل إلى «k» حيثُ تكون ترجمة «$e_{i1} \ldots. e_{ik}$» باحتمالات: $P(e_{i1} \mid f_j) \ldots. \ldots. P(e_{ik} \mid f_j)$.

٤-متوسط عدد ترجمات كل كلمة هو «k».

٥-كل كلمة تترجم إلى كلمة واحدة.

٦-كل كلمة تترجم في نفس المكان في الجملة المترجمة.

لو افترضنا كل هذه الافتراضات الغير واقعية فإنَّ عدد ترجمات الجملة «f» يكون «k^m». فعلى سبيل المثال إذا كان عدد كلمات الجملة «f» عشر كلمات (m=10)، وكل كلمة يمكن أن تُترجم إلى كلمتين مختلفتين في المتوسط (k=2)، فإن عدد الجمل التي يمكن أن تنتج هو 2^10؛ أي ١٠٢٤ ترجمة بـ ١٠٢٤ احتمال. ويكون احتمال ترجمة الجملة «f» إلى الجملة «e» كما يلي:

$$P(e \mid f) = \prod_{j=1}^{m} P(e_j \mid f_j) \qquad (٢)$$

بعد القيام بحساب ١٠٢٤ احتمال، نختار الجملة الأكثر احتمالا.

مُشكلةُ هذه الطريقة أنَّ الترجمة تعتمد فقط على احتمالات ترجمة الكلمات التي ينبغي أن تكون جيِّدة جدًّا حتى يمكن الحصول على ترجمة مقبولة. في الواقع يصعب الحصول على تقدير جيد لترجمة كل الكلمات من لغة إلى لغة أخرى؛ لذلك تم استخدام قاعدة بايز (Bayes' Rule) كما هو موضح في المعادلة رقم (٣)

$$P(e \mid f) = P(f \mid e)\ P(e) / P(f) \qquad (٣)$$

هذه المعادلة تحول حساب احتمال ترجمة جملة من لغة المصدر إلى لغة الهدف (P(e | f)) إلى حساب احتمالين، الاحتمال الأول هو احتمال ترجمة جُملة من لغة الهدف إلى لغة المصدر (P(f | e)) والاحتمال الآخَر هو احتمال حدوث هذه الجملة في لغة الهدف (P(e)). أما احتمال حدوث جملة لغة المصدر (P(f)) فهو قيمة ثابتة يمكن حذفها للتبسيط، وبذلك تصبح المعادلة رقم (٣) كما هو موضح في المعادلة رقم (٤)

$$P(e \mid f) = P(f \mid e)\ P(e) \qquad (٤)$$

الاحتمال الأول يمكن حسابه من نموذج الترجمة والاحتمال الآخَر يمكن حسابه من نموذج اللغة. في أول نموذج للترجمة قدمه مركز أبحاث "IBM - IBM Model-1" كان نموذج الترجمة مكونًا من مجموعة احتمالات لترجمة كلمات من لغة الهدف إلى لغة

المصدر، يتم حسابها من مدونة ثنائية اللغة. أما نموذج اللغة فهناك نماذج عديدة للغة، أبسطها هو نموذج اللغة الثنائيّ، وهو عبارة عن مجموعة من احتمالات تتابع كلمة لكلمة أخرى؛ ويمكن تكوين هذا النموذج من مدونة أحادية اللغة.

ومن ثم يمكن حساب احتمال ترجمة جملة مكونة من عدة كلمات من خلال حساب المعادلة (٤) كما يلي:

$$P(f\,|\,e) = \prod_{j=1}^{m} P(f_j\,|\,e_j)\ \ P\,P(e_j\,|\,e_{j-1}) \quad \text{with } P \text{ over } j=1 \ldots m+1 \qquad (٥)$$

هذه الطريقة أفضل من الطريقة المباشرة، حيث إنَّ نموذج اللغة يحسن من جودة الترجمة لأنه يعطي وزنا أكثر للترجمة التي تتوافق مع قواعد لغة الهدف، ومن ثم فإن المشكلة تصبح في كيفية بناء نموذج الترجمة ونموذج اللغة.

لبناء نموذج الترجمة سنكون في حاجة لتكوين مدونة متحاذية ثنائية اللغة، وهذه أيضا مشكلة حيث إنَّ المدونات اللُّغَوِيَّة الثنائية لا تكون متحاذية على مستوى الكلمة حين يتم تجميعها؛ ومحاذاة المدونة الثنائية على مستوى الكلمة يدويًّا فيه صعوبة بالغة نظرًا لكبر حجم المدونات الثنائية التي تستخدم لإنتاج نموذج ترجمة جيد. لذلك فلإنتاج نموذج ترجمة يتمّ اتِّباع الخطوات التَّالية القائمة على فكرة خوارزم التقدير والتعظيم (Estimation- Maximization Algorithm):

١- يتم توليد جميع المحاذات الممكنة على مستوى الكلمة لكل جملتين متقابلتين.

٢- يتم حساب احتمال ترجمة كل كلمة من لغة الهدف إلى لغة المصدر P(f | e) تقريبيًّا عن طريق افتراض أن ترجمة أي كلمة في لغة الهدف يمكن أن تكون واحدة من الكلمات في لغة المصدر؛ وإذا كان عدد الكلمات في لغة المصدر هو (N) فسيمكن حساب احتمال P(f | e) تقريبيًّا كالتَّالي:

$$P(f\,|\,e) = 1/N \qquad (٦)$$

٣- يتم حساب احتمال كل محاذاة من خلال المعادلة الآتية:

$$P(a, f\,|\,e) = \prod_{J=1}^{m} P(f_j\,|\,e_{a_j}) \qquad (٧)$$

حيث «a» هو كمّ متجه (Vector) يمثل المحاذاة بين الكلمات في جملة المصدر والكلمات في جملة الهدف. فعلى سبيل المثال، إذا كانت الجملة الإنجليزية :

(A Chinese minister wins United Nations environmental prize)

وترجمتها إلى اللغة العربية:

«وزير صيني يفوز بجائزة الأمم المتحدة للبيئة»

ومع اعتبار أن لغة المصدر هي الإنجليزية ولغة الهدف هي العربية، حيثُ إنَّ كمّ المتجه «a» كالتالي:

٤	٧	٥	٦	٣	١	٢	
Prize	Environmental	Nations	United	Wins	Minister	Chinese	A

فان «f٥» في المعادلة رقم (٧) هي كلمة «United» و «j a e» في نفس المعادلة هي «e a٥» حيث «a٥» تعني الرقم ٦ الذي يمثل الكلمة السادسة في جملة الهدف، وهي الجملة العربية؛ ومن ثم يمثل «j a e» كلمة «المتحدة». ويكون احتمال هذه المحاذاة:

(٨) P (a, A Chinese minister… | … وزير صيني) = P (A |null) P (Chinese| صيني).....

لكل محاذاة احتمال، وهذا الاحتمال يتناسب طرديًّا مع تحسُّن احتمالات ترجمة الكلمات المتحاذية. في البداية تكون جميع احتمالات توليد كلمات جملة المصدر من جملة الهدف متساوية كما سبق وتم شرحه في الخطوة رقم (٢). بعد حساب الاحتمالات لكل محاذاة لجملتين متقابلتين يتم تطبيع (normalize) هذه الاحتمالات ليكون مجموعها ١,٠ .

٤- من خلال احتمالات المحاذاة المختلفة لكل جملتين متقابلتين، فإنه يمكن أن يتم إعادة حساب نموذج الترجمة، والذي يتكون من مجموعة من احتمالات توليد كلمات من لغة الهدف إلى لغة المصدر من خلال القيام بعملية عدٍّ جزئيّ (partial count, pc) طبقًا للمعادلة التالية:

$$pc(f \mid e) = \sum_{a} pc\,(a\,,\, f \mid e) \qquad (٩)$$

والعد الجزئيّ له علاقة باحتمال المحاذاة. فعلى سبيل المثال، إذا كانت كلمة «للبيئة» قد تمت محاذاتها بكلمة (environmental) في جملتين متقابلتين وكان احتمال محاذاة هاتين الجملتين «a» هو ٤,٠، كان العد الجزئي «pc» لترجمة كلمة «للبيئة» إلى (environmental)

هو ٠,٤. إذا كانت هاتان الكلمتان قد تمت محاذاتهما في جملتين أخريين وكان احتمال المحاذاة هو ٠,٣، فإن العد الجزئي لتوليد كلمة (environmental) من كلمة «للبيئة» يصبح ٠,٧؛ وهكذا تتم زيادة العد الجزئي كلما حدثت محاذاة بين نفس الكلمتين. أما إذا تم توليد كلمة (environment) من كلمة «للبيئة» في عدة جمل وكان العد الجزئي لهذا التوليد هو ١,٥ فإنه يمكن حساب احتمالات توليد كلمة (environment) و(environmental) من كلمة «للبيئة» من خلال قسمة العدّ الجزئيّ لكل حالةٍ على مجموع العدّ الجزئي للحالتين، ومن ثم يكون احتمال توليد كلمة (environmental) من كلمة «للبيئة»:

P (environmental | للبيئة) = ٠,٧ / (٠,٧+١,٥) = ٠,٣٢

واحتمال توليد كلمة (environment) من كلمة «للبيئة»:

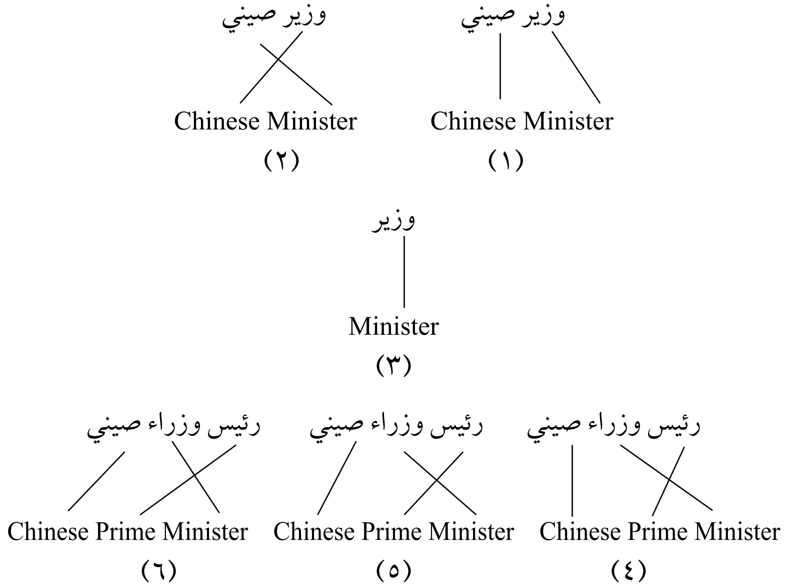
P (environment | للبيئة) = ١,٥ / (٠,٧+١,٥) = ٠,٦٨

٥- يتم إعادة حساب احتمالات المحاذاة لجميع الجمل بعد إعادة حساب احتمالات توليد كلمات لغة المصدر من لغة الهدف، وتتمّ مقارنة هذه الاحتمالات الجديدة مع احتمالات المحاذاة القديمة؛ فإذا كانت نتيجة المقارنة أن هناك فارقًا كبيرًا، فستتم إعادة الخطوة رقم ٤؛ أما إذا كان هذا الفارق صغيرًا جدًّا فسيتم الانتهاء من هذه العملية؛ وتكون نتيجة هذه العملية بناء نموذج الترجمة، وكذلك إنتاج مدونة متحاذية على مستوى الكلمات.

وسوف نعطي هنا مثالاً تطبيقيًّا مبسطًا لتوضيح العملية السابقة. لو افترضنا أننا نملك هذه المدونة:

Chinese Minister	وزير صيني
Minister	وزير
Chinese Prime Minister	رئيس وزراء صيني

وبتطبيق الخطوات السابقة على هذه المدونة نحصل على الآتي:

١. توليد جميع المحازات الممكنة على مستوى الكلمة لكل جملتين أو عبارتين متقابلتين (هناك محازات أخرى، ولكننا سنكتفي بهذه المحازات للتبسيط).	
٢. حساب احتمال ترجمة كل كلمة من لغة الهدف إلى لغة المصدر.	حيث إن عدد كلمات لغة المصدر في المدونة البسيطة هو ثلاث كلمات، فإن أي كلمة في لغة الهدف يمكن أن تولد أيًّا من هذه الكلمات. ويكون احتمال توليد أي كلمة في لغة المصدر هو ١/ ٣.
٣. حساب احتمال كل محازاة.	احتمالات محازات العبارتين المتقابلتين الأوليين في هذه المدونة المبسطة هي: $P(a, f\|e)= 1/3X1/3 = 1/9$ وحتى يكون مجموع احتمالات المحازاة ١، وحيث إن هناك محازاتين لهاتين العبارتين المتقابلتين، فإن احتمال كل محازاة هو ١/ ٢. بالنسبة للكلمتين المتقابلتين في المحازاة رقم (٣)، فإن احتمال هذه المحازاة سوف يكون ١,٠ . بالنسبة للعبارات المتقابلة في المحازات (٤) و (٥) و (٦) فإن احتمال كل محازاة سوف يكون ١/ ٣.

<table>
<tr><td rowspan="11">٤. احتمالات توليد كلمات لغة المصدر من لغة الهدف.</td><td>المحازاة (٤)</td><td>pc(prime | رئيس) = 1/3</td></tr>
<tr><td>المحازاة (٥)</td><td>pc(minister | رئيس) = 1/3</td></tr>
<tr><td>المحازاة (٦)</td><td>pc(Chinese | رئيس) = 1/3</td></tr>
<tr><td>المحازاة (٤)</td><td>pc(minister | وزراء) = 1/3</td></tr>
<tr><td>المحازاة (٥)</td><td>pc(Chinese | وزراء) = 1/3</td></tr>
<tr><td>المحازاة (٦)</td><td>pc(prime | وزراء) = 1/3</td></tr>
<tr><td>المحازاة (١) و (٣)</td><td>pc(minister | وزير)= 1+1/2</td></tr>
<tr><td>المحازاة (٢)</td><td>pc(Chinese | وزير)= 1/2</td></tr>
<tr><td>المحازاة (١) و (٤)</td><td>pc(Chinese | صيني)= 1/2+1/3</td></tr>
<tr><td>المحازاة (٢) و (٦)</td><td>pc(minister | صيني)= 1/2+1/3</td></tr>
<tr><td>المحازاة (٥)</td><td>pc(prime | صيني) = 1/3</td></tr>
<tr><td></td><td colspan="2">من هذا العد الجزئي يمكن إعادة حساب احتمالات توليد كلمات لغة المصدر من لغة الهدف كالتالي:

P(minister | صيني) = (5/6)/(5/6+5/6+1/3) = 5/12
P(Chinese | صيني) = (5/6)/ (5/6+5/6+1/3) = 5/12
P(prime | صيني) = (1/3)/ (5/6+5/6+1/3) = 2/12 = 1/6
P(minister | وزير) = (3/2)/ (3/2+ 1/2) = 3/4
P(Chinese | وزير) = (1/2)/ (3/2+ 1/2) = 1/4
P(prime | رئيس) = (1/3)/(1/3+1/3+1/3) = 1/3
P(minster | رئيس) = (1/3) /(1/3+1/3+1/3) = 1/3
P(Chinese | رئيس) = (1/3) /(1/3+1/3+1/3) = 1/3
P(minister | وزراء) = (1/3) /(1/3+1/3+1/3) = 1/3
P(Chinese | وزراء) = (1/3) /(1/3+1/3+1/3) = 1/3
P(prime | وزراء) = (1/3) /(1/3+1/3+1/3) = 1/3</td></tr>
</table>

٥. إعادة حساب احتمالات المحازاة.	احتمالات المحازاة قبل التطبيع	احتمالات المحازاة المطبعة
	$P(a=1, f\|e) = 3/4X5/12 = 15/48$ $P(a=2, f\|e) = 1/4X5/12 = 1/48$ $P(a=3,f\|e)=3/4$ $P(a=4, f\|e) = 1/3X1/3 X5/12= 5/72$ $P(a=5, f\|e) = 1/3X1/3X1/6 =2/72$ $P(a=6, f\|e) = 1/3X1/3X5/12 = 5/72$	$P(a=1, f\|e) = 15/16$ $P(a=2, f\|e) = 1/16$ $P(a=3, f\|e) = 1$ $P(a=4, f\|e) = 5/12$ $P(a=5, f\|e) = 2/12$ $P(a=6, f\|e) = 5/12$
	كما نرى فإن المحازاة رقم (١) والمحازاة رقم (٤) والمحازاة رقم (٦) قد تحسنت بقدر كبير، وذلك بسبب المحازاة رقم (٣)، والتي تحتوي على كلمة واحدة مما يعطي دفعة كبيرة لاحتمال توليد كلمة minister من كلمة وزير، وكذلك لأن احتمال توليد كلمتي Chinese, minis-ter من كلمة صيني أكبر من احتمال توليد كلمةprime من كلمة صيني. ومن ثم فإننا سوف نعيد الخطوة رقم (٤).	

٦. إعادة احتمالات توليد كلمات لغة المصدر من لغة الهدف.

pc	المحازاة
pc(minister | صيني) = 1/6+5/12 = 7/12	المحازاة (٢) و (٦)
pc(Chinese | صيني) = 15/16+5/12 = 65/48	المحازاة (١) و (٤)
pc(prime | صيني) = 2/12	المحازاة (٥)
pc(minister | وزير) = 15/16+1 = 31/16	المحازاة (١) و (٣)
pc(Chinese| وزير) = 1/16	المحازاة (٢)
pc(prime |رئيس) = 5/12	المحازاة (٤)
pc(minister| رئيس) = 2/12	المحازاة (٥)
pc(Chinese | رئيس) = 5/12	المحازاة (٦)
pc(minister | وزراء) = 5/12	المحازاة (٤)
pc(Chinese| وزراء) = 2/12	المحازاة (٥)
pc(prime| وزراء) = 5/12	المحازاة (٦)

من هذا العد الجزئي يمكن إعادة حساب احتمالات توليد كلمات لغة المصدر من لغة الهدف كالتالي:

P(minister | صيني) = (7/12)/(7/12+65/48+2/12) = 28/101

P(Chinese | صيني) = (65/48)/(101/48) = 65/101

P(prime | صيني) = (2/12)/ (101/48) = 8/101

P(minister | وزير) = (31/16)/ (31/16+ 1/16) = 31/32

P(Chinese | وزير) = (1/16)/ (32/16) = 1/32

P(prime | رئيس) = (5/12)/(5/12+2/12+5/12) = 5/12

P(minster | رئيس) = (2/12) /(12/12) = 2/12

P(Chinese | رئيس) = (5/12) /(12/12) = 5/12

P(minister | وزراء) = (5/12) /(5/12+2/12+5/12) = 5/12

P(Chinese |وزراء) = (2/12) /(12/12) = 2/12

P(prime |وزراء) = (5/12) /(12/12) = 5/12

يمكننا أن نلاحظ أن احتمالات توليد كلمات لغة المصدر الصحيحة من لغة الهدف قد تحسنت بشكل ملحوظ.

ببليوجرافيا مرجعيَّة

1. Anastasiou, D. (2010). Idiom Treatment Experiments in Machine Translation. Cambridge Scholars Publishing.
2. Badr, I.; Zbib, R.; Glass, J. (2008). Segmentation for English-to-Arabic Statistical Machine Translation, Proceedings of Proceeding of HLT-NAACL-Short 2008, Stroudsburg, PA, USA, pp. 153-156.
3. Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. In International Conference on Learning Representations (2015).
4. Bhattacharyya, P. (2013): Machine Translation. Taylor & Francis Group.
5. Blokdyk, G. (2018). Statistical Machine Translation a Clear and Concise Reference. Emereo Pty Limited.
6. Brown et al., A Statistical Approach to Machine Translation, Computational Linguistics, 1990.
7. Brown et al., The mathematics of statistical machine translation: parameter estimation. Computational Linguistics 1993.
8. Casacuberta et al. (2014). Francisco Casacuberta, Marcello Federico, and Philipp Koehn, editors. 2014. AMTA 2014 Workshop on Interactive and Adaptive Machine Translation (IAMT 2014), Vancouver, Canada, October. Association for Machine Translation in the Americas (AMTA).
9. Chan, S. (2006): A Dictionary of Translation Technology, Publisher: The Chinese University Press.
10. Chan, S. (2018). The Human Factor in Machine Translation. Routledge.
11. Chiang, D. (2005). “A hierarchical phrase-based model for statistical machine translation”. In Proc. ACL, pages 263-270.

12. Cho, K., van Merrienboer, B., Gülçehre, Ç., Bougares, F., Schwenk, H., and Bengio,Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Conference on Empirical Methods in Natural Language Processing (2014).

13. Clark, J. (2015). Locally non-linear learning via feature induction and structured regularization in statistical machine translation. In Dis-sertation, Carnegie Mellon University.

14. Denkowski et al. (2014a). Learning from post-editing: Online model adaptation for statistical machine translation. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 395-404, Gothenburg, Sweden, Association for Computational Linguistics.

15. Denkowski et al. (2014b). Real time adaptive machine translation for post-editing with cdec and TransCenter. In Proceedings of the EACL 2014 Workshop on Humans and Computer-assisted Translation, pag-es 72-77, Gothenburg, Sweden, Association for Computational Linguistics.

16. Devlin, J., Zbib, R., Huang, Z., Lamar, T., Schwartz, R. M., and Makhoul, J. Fast and robust neural network joint models for statistical machine translation. In ACL (1) (2014), Citeseer, pp. 1370-1380.

17. Ebrahim et al. (2015). English-Arabic Statistical Machine Translation: State of the Art. A. Gelbukh (Ed.): CICLing 2015, Part I, LNCS 9041, pp. 520-533.

18. Galley, M.; Hopkins, M.; Knight, K.; Marcu, D. (2004). “What’s in a Translation Rule?”, Proc. NAACL-HLT.

19. Germann, U. (2014). Dynamic phrase tables for machine translation in an interactive post-editing scenario. In Proceedings of the AMTA 2014 Workshop on Interactive and Adaptive Machine Translation, pages 20-31.

20. Goutte, C.; Cancedda, N.; Dymetman, M.; Foster, G. (2008). Learning Machine Translation (Neural Information Processing series), Publisher: The MIT Press.

21. Green, S. (2014). Mixed-initiative natural language translation. In Dissertation, Stanford University.

22. Habash et al. (2009). MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. In Khalid Choukri and Bente Maegaard, editors, Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, Egypt, April. The MEDAR Con-sortium.

23. Habash, N.; Hu, J. (2009). Improving Arabic-Chinese Statistical Machine Translation using English as Pivot Language, Fourth Workshop on Statistical Machine Translation, EACL 2009, Athens, Greece.

24. Habash, N.; Sadat, F. (2006). Arabic Preprocessing Schemes for Statistical Machine Translation, Proceedings of HLT- NAACL, New York City, New York, USA.

25. Hassan et.al. (2008). Syntactically Lexicalized Phrase-Based SMT, IEEE, Transaction on Audio, Speech, and Language processing.

26. Ittycheriah and Roukos, Direct Translation Model2, NAACL/HLT 2007.

27. Khemakhem, I.T.; Jamoussi, S. (2013): Integrating morpho-syntactic features in English-Arabic statistical machine translation. In: ACL 2013, pp. 74.

28. Koehn, P. (2010): Statistical Machine Translation, Publisher: Cambridge University Press; 1st edition.

29. Koehn, P. (2016). The State of Neural Machine Translation (NMT). Omniscien Technologies. Retrieved 2019-1-31.

30. Lee, Y. S. (2004): Morphological Analysis for Statistical Machine Translation, Proceeding of HLT-NAACL-Short 2004, Stroudsburg, PA, USA, pp. 57-60

31. Matusov et.al., (2008). System Combination for Machine Translation of Spoken and Written Language, IEEE, Transaction on Audio, Speech, and Language processing.

32. Nabahan, A.; Rafea, A. (2010). A Hybrid Noun Phrase Translation System, 7th International Conference on Informatics and Systems (INFOS), vol., no., pp.1-7, 28-30.

33. Nădejde, M. (2018). Syntactic and Semantic Features for Statistical and Neural Machine Translation. University of Edinburgh.

34. N-Varela, C. M.; Bartrina, F. (2013): The Routledge Handbook of Translation Studies. Routledge.

35. Olive, J.; Christianson, C.; McCary, J. (2011). Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation, Publisher: Springer; 1st Edition. Edition.

36. Olive, J. (2011): Handbook of Natural Language Processing and Machine Translation. Springer.

37. Poibeau, T. (2017). Machine Translation. MIT Press.

38. Ragab, A.; Rafea, A., Enhancing Phrase Extraction from Word Alignments Using Morphology, Proceedings of the 5th Conference on Language Engineering, Cairo, Sept. 2005.

39. Ragab, A.; Rafea, A., Tuning Statistical Machine Translation Parameters Using Perplexity, IEEE International Conference on Information Reuse and Integration 2005, IEEE IRI-2005, August 15-17, 2005, Hilton, Las Vegas, Nevada, USA.

40. Rutkowski, S. (2012): Machine Translation Evaluation: An Analysis of Two Translations Produced by Google Translate and English Translator XT. Lambert Academic Publishing.

41. Scott, J. (2019). Legal Translation Outsourced. Oxford University Press.

42. Shaalan, K.; Rafea, A.; Baraka, H. (2003). Proposed Approach for Generating Arabic from Interlingua in a Multilingual Machine Translation System, 4th Conference on Language Engineering, Cairo, Egypt, October 2003.

43. Soudi, A. (2012): Challenges for Arabic Machine Translation. John Benjamins Publishing.

44. Sutskever, I.; Vinyals, O.; and Le, Q. V. Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems (2014), pp. 3104-3112.

45. Wilks, Y. (2010). Machine Translation: Its Scope and Limits, Springer.

46. Wu, D. (2013): Foundations of Text Alignment: Statistical Machine Translation Models from Bitexts to Bigrammars. Springer London.

47. Zetzsche, J. (2017). Translation Matters. CreateSpace Independent Publishing Platform.

48. Zollmann, A.; Venugopal, A.; Och, F.; Ponte, J. “A Systematic Comparison of Phrase-Based, Hierarchical and Syntax-Augmented Statistical MT”, Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 1145-1152, Manchester, August 2008.

الفصل الثَّالث
التَّشكيل الآليّ

د. مُحسن رشوان

١ – تعريف بعلامات التَّشكيل في اللُّغة العربيَّة.
٢ – صياغة رياضيَّة لحسم مشكلة التشكيل.
٣– مصنِّف بايز المبسط.
٤ – خوارزم فيتربي.
٥ – مسائل أخرى متشابهة.
٦– أفضل ما سُجِّلَ من نتائج.
٧– طبيعة الموارِد اللُّغَوِيَّة الَّتي نحتاجها
٨– أفكارٌ بحثيَّة لأطرُوحاتٍ علميَّةٍ مُستَقبليَّة.

تمهيد

الكلمة العربية مركبة تركيبا مميزا؛ فهي تجمع بين خاصية الاشتقاق (Derivative) وخاصية الالتصاق (Adhesion)، ويبدو أن الأمَّة العربيَّةَ تقدمت بلغتها تقدما هائلا حتى أن بلغائها لم يكونوا في حاجة عند كتابتها إلى استخدام النقاط أو علامات الضَّبط (الحركات القصيرة والشَّدَّة والتَّنوين) لبيان المعنى. ورحم الله المتأخرين بأن قيد الله من أضاف التنقيط وترك ما دون ذلك من علامات لبداهة القارئ. وعندما ظهر علم حوسبة اللغة أضاف هذا الأمرُ المزيدَ من التَّحدِّيات أمام حوسبة اللغة العربية مقارنة بلغات أخرى ليس فيها هذا التحدي.

وسنتناول في هذا الفصل سبعة محاور، هي:

١- تعريف بعلامات الضَّبط/ التَّشكيل في اللُّغة العربيَّة.

٢- صياغة رياضية لحسم مشكلة التشكيل.

٣- مصنِّف بايز المبسط (Naïve Bayesian Classifier).

٤- خوارزم فيتربى (Viterbi Algorithm).

٥- مسائل أخرى متشابهة.

٦- أفضل ما سُجِّلَ من نتائج.

٧- طبيعة الموارِد اللُّغَوِيَّة الَّتي نحتاجها.

١- تعريف بعلامات التَّشكيل في اللُّغة العربيَّة

توجد بعض اللغات مثل: اللغة الإنجليزية، غالبا ما يُحَدَّدُ نطق الكلمات بها من خلال الحروف المكونة لها. حيث إنَّ تتابع الحروف المتحركة والساكنة هو الذي يحدد النطق الصحيح للكلمة. ويطلق على مثل هذه اللغات (اللغات غير المشَّكَلة). ومن ناحية أخرى، توجد بعض اللغات تعتبر نطق كلماتها غير محددة بالكامل بواسطة أحرف هجائها فقط. فعلى سبيل المثال: من الممكن أن تكون هناك كلمتان متطابقتان في التهجئة ولكنهما مختلفتان في النطق والمعنى تمامًا. لإزالة ذلك الالتباس يتم وضع علامات خاصة

بأعلى أو أسفل الكلمة لتحديد النطق الصحيح. وتلك العلامات تسمى «التشكيل»، واللغات التي تستخدم تلك العلامات تسمى اللغات المشَّكَلة. واللغة العربية واحدةٌ من هذه اللغات. الواقع أنَّ اللغة العربية لديها أدقّ نظام تشكيل مفصل.

ويبين الجدول التَّالي علامات التشكيل في اللغة العربية ومغزى كلٍ منها:

التشكيل		العلامة	أمثلة	ملاحظات
١	عَ	فتحة	مكَافَأة، مَصنَع، بَرَاءَة، عَلَم	-
٢	عُ	ضَمَّة	كُتُب، هُمُوم، صُراخ، عُود	-
٣	عِ	كسرة	كِتاب، مِهنة، عِيال، هِمم	-
٤	عْ	سكون	عوْن، إنْسان، رأْي، استعْمال	-
٥	عًا	تنوين فتحة	كتابًا، نهايةً، طعامًا، ثراءً	الأصل أن يوضَع تنوين الفتح على الحرف السَّابق للألف
٦	عٌ	تنوين ضمة	حصْرٌ، قصورٌ، استعدادٌ، سردٌ	يحتوي الحرف الأخير فقط على التشكيل
٧	عٍ	تنوين كسرة	مساءٍ، ملاقاةٍ، معانٍ، محامٍ	يحتوي الحرف الأخير فقط على التشكيل
٨	ا، و، ى	مَدّ	كَاتب، مغَانم، قَال، حِمِير، عِيد، طِين، بيُوت، كُوفيء، رُوح	-
٩	ى	ألِف لَيِّنة	مصطفَى، مَشْتَى، نادَى، مغالَى	يحتوي حرف الياء فقط على التشكيل
١٠	ا	حرف غير منطوق	اَلسَّماء، والسَّماء، قالُوا، أُولئك	-
١١	ع	مَدٌّ مستتر بالألف	هذا، ذلك، الرحمن، هؤلاء	عادة لا يكتب هذا الصنف من التشكيل

التشكيل		العلامة	أمثلة	ملاحظات
١٢	عّ	شَدَّة	مُعَلِّم؛ لِّ = لْ+لِ كُتَّاب؛ تَّ = تْ+تَ حقٌّ؛ قٌّ = قْ+قٌ الصُّبح؛ صُّ = صْ+صُ	في الحقيقة، لا تعتبر الشدة من التشكيل ولكنها تعتبر علامة فقط لتوضح مضاعفة الحرف عند النطق

الجدول ٣-١ : علامات التشكيل في اللغة العربية.

وينبغي أن يحتوي كل حرف في الكلمة العربية على سمتين عند تشكيله، تُمَثِّلان معلومات الحرف المشكل، هما:

١,١- حالة الحرف المُشَدَّد (يحتوي على شدة / لا يحتوي عليها).

١,٢- الحرف المشكل.

مع الأسف، فإن كتابة اللغة العربية لم تعد تتضمن علامات التشكيل. فقد استعاض الناس عن التشكيل بمعرفتهم بالنطق الصحيح من خلال السياق، وأصبح التشكيل يستخدم فقط لإزالة الالتباس في بعض المواضع أو لأغراض تعليمية. ولهذا فإن المشكل الآلي يجب أن يتدرب على تشكيل الكلمات العربية ويتضمنَ آلية للتعرف على أي علامات تشكيل ناقصة بالكلمة العربية المدخلة.

ونختم هذا المحور بتعريف حالات التشكيل الثلاثة المختلفة في الكلمة العربية:-

التشكيل التام:

حيث يتم تحديد كافة المعلومات التشكيلية في اللغة العربية لكل حرف في الكلمة، متضمنة الحرف الأخير، وأحيانا يتم تشكيل الحرف الأخير اعتمادًا على التحليل النحوي للكلمة؛ ويتم ذلك من سياق الجملة. انظر هذا المثال:

لا يوجد تشكيل: إذا كنت ذا قلب قنوع فأنت ومالك الدنيا سواء

تشكيل جزئي: إذا كنتَ ذا قلبٍ قنوعٍ فأنت ومالِك الدٌنيا سَوَاءُ

تشكيل تام: إذَا كُنْتَ ذَا قَلْبٍ قَنُوعٍ فَأَنْتَ وَمَالِكُ الدٌنْيَا سَوَاءُ

٢- صياغة رياضيَّة لحسم مشكلة التشكيل

دعنا نأخذ مثالا مبسطًا لفهم المسألة رياضيا؛ سنرمز للكلمة بالرمز اللاتيني (اختصارًا لـ word). ويسأل سائلٌ: لماذا تكون المعادلات بالأحرف اللاتينية؟ إن ذلك لوصل القارئ بمعارف العصر؛ فهذا الكتاب نريده أن يكون همزة وصل بعلوم ومراجع كثيرة، الأجنبي فيها أكثر من العربي بآلاف المرات؛ فلا ضير من ذلك؛ بل إنَّ فيه نفعَ التَّعوُّد على الانتفاع من علوم سُبقنا فيها أجيالا. لعل أجيالاً قادمة يتدفق منها عطاء أهل العربية من العلوم الحديثة ما يرجح كفة الترميز بها والكتابة بها ليعود النهل منها كما كان من قبل.

فلو افتَرضنا أن الجملة تتكون من العديد من الكلمات كالآتي:

$$w_1\ w_2 \ldots w_{(j-1)}\ w(j)\ w_{(j+1)} \ldots w_{(N-1)}\ w_N$$

فهذه الجملة التي عدد كلماتها N ورقم الكلمة في الجملة «j» و «...» تعنى أن هناك كلمات لها أرقام متصاعدة من آخر كلمة قبل هذه النقط إلى أول كلمة بعدها؛ سيكون لكل كلمة أكثر من تشكيل محتمل إذا أُخذت مجردة عن سياق الجملة. وكمثال على ذلك:

١- التلميذ كتب الدرس ← كَتَبَ

٢- التلميذ حمل كتب المدرسة ← كُتُبَ

«كتب» يمكن أن تأخذ تشكيلات كثيرة ولكل معنى مختلف مثال:

كَتَبَ، كُتُبٌ، كُتِبَ،......

في بعض الأحيان تكون الأشكال الصحيحة المختلفة لتشكيلات الكلمة بالعشرات. ولتيسير ذلك.. نفرض أن عندنا عددا محدودا من الاحتمالات لتشكيل كلمة بينها، ولتكن كلمة ونفرض أن لها M من الحلول (Solutions)؛ وتعال نسمي هذه الحلول: الحل الأول S_1 والحل الثاني S_2،... وهكذا.

وتعال أيضا نرمز للسياق (Context) بالرمز C، ويمكننا الاصطلاح علي أن السياق هو باقي كلمات الجملة كلها أو أن نحدد هذا السياق بعدد محدود من الكلمات قبل وبعد

هذه الكلمة، هذا كله جائز. تعال نُسغ المعادلة الرياضية لاحتمال كل حل:

١- احتمال الحل ١: s_1، باعتبار السياق C هو $P(s_1/C)$

٢- احتمال الحل ٢: s_2، باعتبار السياق C هو $P(s_2/C)$

٣- احتمال الحل ٣: s_3، باعتبار السياق C هو $P(s_3/C)$

وهكذا، سواء أكان هناك حلان أو أكثر فإننا نفاضل بين هذه الحلول ونختار الأعلى احتمالا. ولكن كيف يمكن لنا أن نحسب هذه الاحتمالات؟ الذي نملكه هو ذخيرة لغوية للتدرب عليها، بها العديد من المرات التي مر بها كل حل من حلول كلمة w_j. والذي نملكه أيضا هو عزل المرات التي جاء فيها كل حل، وعندئذ تكون حساباتنا لـ $P(C/s_j)$ وليس $P(s_j/C)$ (حيث j هنا تشير إلى رقم الحل)، وذلك لأننا نعزل الجمل التي مر بها الحل لا الجمل التي جاءت بالسياق C، وهنا لا بد من اللجوء لمعادلة بايز (Bayes):

$$P(s_j/C) = \frac{P(C/s_j)\ P(s_j)}{P\ (C)} \qquad (1)$$

ولما كان المقام ثابتا لكل الحلول، فلا داعي لحسابه، ويكفي أن نحسب بسط يمين المعادلة (١). وطالما أن البسط للحل الصحيح هو الأعلى فإن ذلك يعني أن هذا الحل هو الأكثر احتمالا. ولأننا لو حذفنا المقام فلن يمثل الطرف الأيسر احتمالا - لكنه يتناسب مع الاحتمال؛ فسوف نعيد صياغة المعادلة مع إعادة تسمية $P(s_j/C) \leftarrow g(s_j/C)$.

$$g(s_j/C) = P(s_j/C) \qquad (2)$$

ويبقى الأمر كما هو، أنه كلما زاد احتمال أي حل زادت قيمة $g(s_j/C)$ لهذا الحل. كيف نحل هذا النوع من المسائل رياضيا؟ سنجد إجابة هذا السؤال فيما يلي.

٣- مصنِّف بايز المبسط (Classifier Bayesian Naïve)

الواقع أنَّ هناك العديد من الطرق والخوارزمات الرياضية لحل هذا النوع من المسائل وتسمى هذه الخوارزمات «المصَنِّفات» Classifiers، ويحتاج شرحها بالتفصيل إلى كتاب مفصل، ولكننا هنا اخترنا بعضًا من هذه المصنفات، وسنبدأ بمصنف يعد

بسيطا ولكنه فعال ونتائجه لا بأس بها، ويسمى «مصنف بايز المبسط» (Bayes Naïve Classifier)، وجريا على عرف المحترفين من الكتاب عند استخدام مصطلح كثير الاستخدام أن يختصروا اسمه باستخدام الأحرف الأولى، أي (م ب م) ويختصرونه بالإنجليزية أيضًا (NBC). ويسمى المبسط لأن هناك فرضية رياضية لتبسيط الحل وهي اعتبار أن الكلمات التي تمثل السياق مستقلة بعضها عن بعض - وإن كان ذلك في الحقيقة غير صحيح لأن بعض الكلمات يقترن كثيرا بكلمات أخرى - وهذا الفرض سمــح لنا بإمكانية التعامل مع السياق بشكل مبسط. والسياق هو مجموع الكلمات التي سبقت الكلمة مباشرة أو تلتها. ويجوز لنا بهذا الفرض أن نكتب سياق الكلمة W_j كالآتي:

$$P(C)=P(w_1) * P(w_2)\ldots P(w_{j-1}))*P(w_{j+1}))\ldots P(w_N) \quad (3)$$

وكذلك يمكن إعادة كتابة المعادلة (٢) كالآتي:

$$g(s_j/C) =[P(w_1/s_j) * P(w_2/s_j)\ldots P(w_{j-1}/s_j) *P(w_{j+1}/s_j)\ldots P(w_N)] * P(s_j) \quad (4)$$

إن صياغة المعادلة (٤) يجعل الحل في متناول أيدينا. فلو أننا تمكنا من حساب الكميات $P(w_k/s_j)$ $(j=1,\ldots,N)$، ثم حسبنا أيضًا $P(s_j)$ نكون قد حسمنا الأمر كله وعرفنا أي الحلول في هذا السياق هو الأوفق. إن حساب هذه الكميات يمكن الرجوع إليه في ملحق-١ لنظرية الاحتمالات وكذلك فصل «نمذجة اللغة» . ولا يفوتنا هنا أن نذكر بأن الاحتمال $P(s_j)$ يسمى النحو الأحادي، وهو احتمال أن تأتي الكلمة بهذا الحل عموما، بصرف النظر عن السياقات المختلفة (أي: احتمال وجودها ككلمة مفردة). ومثال ذلك: كلمة «قال» من مادة القول قد يصل نحوها الأُحادي - مشروطا بورود كلمة «قال» - إلى أكثر من ٩٩٩, ٠ بينما كلمة «قال» من مادة قيل (أي النوم بالظهيرة) قد لا يصل نحوها الأحادي - مشروطا بورود كلمة «قال» - إلى ٠٠١, ٠.

والجدير بالذكر أننا سوف نقابل عند تطبيق هذا الخوارزم أو هذا المصنف مشكلة وهي أن بعض الكلمات لم نرها من قبل في الذخيرة اللغوية التي تدرب النظام عليها. وفي سياق جديد إذا أتت كلمة واحدة لم تُر من قبل، فسيكون احتمال ورودها صفرا،

وسوف نضرب في صفر فتكون النتيجة صفرا مهما كانت قوة شواهد الكلمات الأخرى في السياق. ولقد واجهنا هذه المشكلة في فصل نمذجة اللغة واستطعنا أن نمنع هذا الصفر بافتراض نسبة احتمال صغيرة نسبيا لما لم نره من الكلمات.

٤- خوارزم فيتربي (Algorithm Viterbi)

وهناك خوارزمات أو مصنفات أخرى مشهورة في حسم مثل هذه المسائل. منها مُصَنِّف فيتربي (Viterbi search) ومصَنِّف (بحث *) أو (A* search) ويعتمد هذان المصنفان على النحو الإحصائي [١٤، ١٥]. وعادة يبحثان عن أفضل مسار عبر الجملة بالكامل أو جزء منها. ولكي نتخيل كيف تعمل هذه المصنفات تعالوا نأخذ مثالا مبسطًا:

ذهب الولد للمدرسة

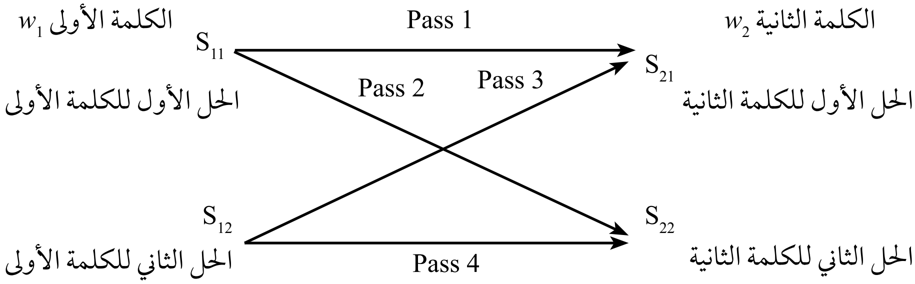
بالتحليل الصرفي لهذه الجملة البسيطة يمكن أن نجد لكل كلمة عددًا من الحلول المحتملة:

	ذهب	الولد	للمدرسة
حل١	ذَهَبَ	الوَلَد	للمدْرَسَةِ
حل٢	ذُهِبَ	الوِلد	للمدَرِّسَةِ
حل٣	ذَهَبٌ		

لننظر إلى كافة الحلول الموجودة (وقد تكون هناك حلول أكثر لبعض الكلمات - وبالتالي للجملة ككل ولكننا سنكتفي بهذه الحلول حتى نتمكن من تتبع المثال). إذا تأملنا كل المسارات الممكنة من الكلمة الأولى إلى الكلمة الأخيرة سوف نجد أننا أمام ١٢ مسارًا (هذا لو اكتفينا بالحلول المبينة في الجدول فقط) ولكل مسار يمكن حساب احتماله باستدعاء حسابات النحو العددي للكلمات. وتهدف هذه الأنواع من المصنفات إلى حساب أفضل هذه المسارات، أي أعلاها احتمالا. بالطبع هي أكثر تعقيدا، ولكن عادة تعطى نتيجة أفضل من المصنف الأول (م ب م) أو NBC. كما يمكن حساب عدد هذه المسارات كالآتي: عدد حلول الكلمة الأولى * عدد حلول الكلمة الثانية * عدد حلول الكلمة الثالثة. وفي حالتنا = ٣ * ٢ * ٢ = ١٢ حلا، أو إن شئت قُل: مسارًا.

وتقدم طريقة أو خوارزم فيتربى حلاً لنفس المشكلة، ولكن بدون فرضية استقلال كلمات الجملة بعضها عن بعض، ولذلك فنتائجها في الغالب أفضل من نتائج خوارزم (م ب م). ولتيسير فهم خوارزم فيتربي (Viterbi) دعنا نأخذ مثالا مبسطا:

هب أن عندنا جملة من كلمتين فقط ولكل كلمة حلاَّن.



الشَّكل ٣-١: المسارات الممكنة لجملة من كلمتين، ولكل كلمة حلاَّن.

ونحتاج إلى حساب احتمال المسارات.

المسار الأوَّل = **Pass1**، المسار الثاني = **Pass2**، المسار الثَّالث = **Pass3**، المسار الرَّابع = **Pass4**

بافتراض أن «Passi» تعنى احتمال المسار i. فإن احتمال المسارات المختلفة:

$$Pass1 = P(s_{11}) * P(s_{11}/\text{بداية الجملة}) * P(s_{21}) * P(s_{21}/s_{11})$$

$$Pass2 = P(s_{11}) * P(s_{11}/\text{بداية الجملة}) * P(s_{22}) * P(s_{22}/s_{11})$$

$$Pass3 = P(s_{12}) * P(s_{12}/\text{بداية الجملة}) * P(s_{21}) * P(s_{21}/s_{12})$$

$$Pass4 = P(s_{12}) * P(s_{12}/\text{بداية الجملة}) * P(s_{22}) * P(s_{22}/s_{12})$$

وأي المسارات أعلى احتمالا نختار التشكيلات عليه لتكون هي الحل.

ولو نظرت ودققت في الحسابات للمسارات الأربعة، ستجد أننا نكرر حساب أجزاء ليست قليلة مع حساب أجزاء تم حسابها في مسارات أخرى؛ فلو أننا استخدمنا ما حسبناه من قبل، يمكن أن نوفر جهدا كبيرا. وحتى تتصور عدد المسارات الممكنة، فلو لدينا جملة بها ٢٥ كلمة (وهو عدد قريب من متوسط عدد كلمات الجملة العربية) ولكل كلمة ثلاثة حلول (بافتراض تساوي عددها لسهولة حسابها - إذ في الواقع يمكن

أن يكون للكلمة حل واحد أو ربما عشرات الحلول)؛ وبالتَّالي يكون عدد المسارات (يُساوي حاصل ضرب (عدد حلول الكلمة الأولى) (عدد حلول الكلمة الثانية) (عدد حلول الكلمة الثالثة)... وهكذا نضرب ٣ في نفسِها ٢٥ مرة)، وعليه، ستكون 3^{25} النتيجة أكثر من ٨٧٤ مليار حل أو مسار محتمل.

أي أننا في حاجة لحساب احتمالات لأكثر من ٨٧٤ مليار مسار محتمل لجملة واحدة؛ ولو افترضنا أنَّ قوة الحاسب ستسمح بحساب احتمال المسار في ١ مللي ثانية، فسنحتاج إلى أكثر من ٨٧٤ مليون ثانية أي حوالي ٨,٢٦ سنة من الحساب المتصل.

تعال نفرض أن قوة الحاسب ءتضاعفت ألف مرة. سنختصر الحل في ٢٣٥ ساعة، أي حوالي عشرة أيام فقط! لن يكون هذا الحل عمليا لولا طرق سريعة لحساب أفضل المسارات، ربما لا تتجاوز جزءًا من الثانية الواحدة بحساب هذه الأيام.

لا بد من التنويه هنا أننا لو لم نتتبع بسهولة بقية الحل لهذا الخوارزم فإن ذلك لن يقلل من فهمنا لروح الحل الذي أسلفناه.

تعال نأخذ المسألة أَعمق قليلا، وهي ٣ كلمات مع حلَّين لكل كلمة.

w_1	w_2	w_3
s_{11}	s_{21}	s_{31}
s_{12}	s_{22}	s_{32}

ويكون لدينا إذن ثمانية مسارات كالآتي:

$$Pass1 = (s_{11}, s_{21}, s_{31})$$
$$Pass2 = (s_{11}, s_{21}, s_{32})$$
$$Pass3 = (s_{11}, s_{22}, s_{31})$$
$$Pass4 = (s_{11}, s_{22}, s_{32})$$
$$Pass5 = (s_{12}, s_{21}, s_{31})$$
$$Pass6 = (s_{12}, s_{21}, s_{32})$$
$$Pass7 = (s_{12}, s_{22}, s_{31})$$
$$Pass8 = (s_{12}, s_{22}, s_{32})$$

وفي هذه الحالة يكون حساب المسار الأول (كمثال لبقيَّة المسارات) كالآتي:

$$Pass1=P(s_{11}\ )*P(s_{11}/\text{بداية الجملة})*P(s_{21)}*P(s_{21}/s_{11},\text{بداية الجملة})$$
$$*P(s_{31})*P(s_{31}/s_{21},s_{11},\text{بداية الجملة})$$

هذا يعطي حسابا دقيقا لاحتمال المسار الأول، ولكنه يتطلب حساب وتخزين مسبق لاحتمالات من النحو الثلاثي 3-gram مثل (بداية الجملة ، $P(s_{21}/s_{11}$ ونحو رباعي مثل (بداية الجملة ، $P(s_{31}/s_{21},s_{11}$ وبالرغم من أن هذا يعطى حلا أكثرَ دقَّة، إلا أن حسابه يتطلب ذاكرة أكبر وحسابات أكثر بكثير من النحو الثنائي. ولذلك يمكن تقريب الحل وتيسير حسابه والحفاظ على الذاكرة المطلوبة في حجم مناسب إذا اكتفينا بالنحو الثنائي (2-gram). ويقرب الحل كالآتي:

$$Pass1\approx P(s_{11}\ )*P(s_{11}/\qquad\qquad)*P(s_{21})$$
$$*P(s_{21}/s_{11})\ *P(s_{31})\ *P(s_{31}/s_{21})$$

عند حساب المسار الأول Pass1 حسبنا المسار الجزئي (s_{11}/s_{21})، فهل عند حساب المسار الثاني Pass2 سنكون في حاجة لإعادة حساب هذا الجزء مرةً أخرى؟ إن Pass2 و Pass1 يشتركان في مسار جزئي وهو(s_{11}/s_{21})، إذن لا داعي لإعادة حسابه ثانية. ويتكرر الموقف بين المسارين Pass3 و Pass4 في حساب (s_{11}/s_{21})، إذ لا داعي لحساب المسار مرتين، وكذلك بين Pass5 و Pass6 فإن حساب(s_{12}/s_{22}) مرة واحدة يكفي. ونفس الشيء بين Pass7 و Pass8 فإن حساب(s_{12}/s_{22}) يكفي مرة واحدة.

ولو أردنا ألا نكرر ما سلف وحسبناه، ثم أضفنا إلى ذلك معلومة أخرى مهمة، هي أننا إذا استطعنا عند أي نقطة أن نحسب أفضل المسارات إليها فلسنا في حاجة للبدء من أول كلمة في كل مرة، بل يكفي أن نرجع للعمود السابق فقط لنكمل الحل عمودا بعد عمود ونحن نتحرك من اليسار إلى اليمين (الواقع أنَّ الكلمات العربية تتحرك من اليمين لليسار، ولكن لأن هذه المسائل موجودة بالكتب الأجنبية من اليسار لليمين؛ وحتى يسهل على القارئ الحل إذا نظر في هذه الكتب فإننا تحركنا في نفس الاتجاه، وهذا لن يغير الحل في شيء). يمكن أن نحسب الآن المسارات جزءًا جزءًا (بدءًا من اليسار إلى اليمين)، وما حسبناه من أجزاء المسار ونحن نتحرك عليه من الكلمة الأولى إلى الكلمة الأخيرة لنتأكد باستمرار أننا نحسب أعلى المسارات احتمالا.

ولسهولة تتبع الحل تعالوا نرمز إلى أفضل الحلول حتى النقطة التي نقف عندها بالرمز L_{ij} ، والذي يعني أن عند هذه النقطة سنجد احتمال أفضل مسار من أول كلمة حتى النقطة ij هو L_{ij} .

يمكن شرح الخوارزم المسمي «فيتربي» كالآتي:

٤ , ١- نبدأ بحساب احتمالات الوصول من حلول الكلمة الأولى إلى حلول الكلمة الثانية.

- نحسب احتمالات الحل الأول للكلمة الأولى مع الحل الأول للكلمة الثانية. ثم احتمال الحل الثاني للكلمة الأولى مع الحل الأول للكلمة الثانية ونأخذ أعلى الحلول احتمالا ونسميه L_{21} . والرمز L_{21} يعنى أفضل الحلول عند الموضع ٢١ أي الصف الأول والعمود الثاني.
- نكرر مع الحل الثاني للكلمة الثانية لنصل إلى L_{22} .

٤ , ٢- نعيد الكَرَّةَ مرةً أخرى، مع الأخذ في الاعتبار أننا عند حساب أفضل المسارات من الأول إلى الآخر سنستفيد من الحسابات السابقة، فلا نبدأ دائما من الأول، ولكن نبدأ من الكلمة السابقة فقط لأننا قمنا بما يلزم قبل ذلك من حساب أفضل المسارات احتمالا حتى هذه الكلمة.

٤ , ٣- تكون القيمة الأعلى بين L_{32} و L_{31} هي احتمال المسار الأعلى احتمالا.

٤ , ٤- ويمكن معرفة المسار (أي أفضل الحلول للكلمات) بالاحتفاظ عند كل خطوة بأفضل المسارات التي انتهينا إليها عند هذه الخطوة.

لا ينقصنا الآن إلا مثالٌ عمليٌّ لتوضيح المسألة. لنأخذ هذا المثال:

«ذهب علي بالكرة» . تعال نفرض أن لكل كلمة حلَّين فقط لتيسير فهم حل المسألة

«ذهب» نفرض لها حَلَّين:

ذَهَبَ (أي تحرك، فعل ماض، حل1 ← s_{11})

ذَهَبٌ (معدن، اسم، حل2 ← s_{21})

«على» نفرض لها حلان:

عَلَى (حرف جر، حل 1 ← s_{21})

عَلي (اسم علم، حل 2 ← s_{22})

«بالكرة» نفرض لها حَلّين:

بالكَرَّة (كرّة بمعنى مرة، حل 1 ← s_{31})

بالكُرَة (كرة يلعب بها، حل 2 ← s_{32})

وتعال نفرض توافر هذه الاحتمالات من مدونة (في حالتنا تخيلية) حسبنا منها الاحتمالات الآتية:

النحو الأحادي 1-gram أو uni-gram

$P(s_{31}) = 0.01$,	$P(s_{21}) =0.05$,	$P(s_{11}) =0.03$,
$P(s_{32}) =0.05$,	$P(s_{22}) =0.01$,	$P(s_{12}) =0.02$,

النحو الثنائي 2-gram أو bi-gram

$P(s_{31}/s_{21})\simeq 0.0$ $P(s_{31}/s_{22}) = 0.05$	P(s_31/s_21) $P(s_{22}/s_{11}) = 0.2$	P(s_11/الجملة بداية) = 0.1
$P(s_{32}/s_{22})\simeq 0.0$ $P(s_{32}/s_{22}) = 0.05$	$P(s_{21}/s_{12})=0.1$ $P(s_{22}/s_{12})=0.1$	$P(s_{12}$/الجملة بداية) = 0.05

الحل الأول للكلمة الأولى s_{11}=ذَهَبَ	الحل الأول للكلمة الثانية عَلَى = s_{21}	الحل الأول للكلمة الثالثة بالكَرَّة = s_{31}
أفضل الحلول حتى هذه النقطة L_{11}=P(s_{11}/ بداية الجملة) *P(s_{11}) = 0.1* 0.03 = 0.003	أفضل الحلول حتى هذه النقطة $L_{12} = max\begin{cases} L_{11} * P(s_{21}/s_{11}) * P(s_{21}) \\ L_{12} * P(s_{21}/s_{12}) * P(s_{21}) \end{cases}$ = $max\begin{cases} 0.003 * 0.05 * 0.05 = 7.5 * 10^{-6} \\ 0.001 * 0.1 * 0.05 = 5 * 10^{-6} \end{cases}$ $= 7.5 * 10^{-6} \rightarrow (s_{11}, s_{12})$	أفضل الحلول حتى هذه النقطة $L_{31} = max\begin{cases} L_{21} * P(s_{31}/s_{21}) * P(s_{31}) \\ L_{22} * P(s_{31}/s_{22}) * P(s_{31}) \end{cases}$ = $max\begin{cases} 7.5 * 10^{-6} * 0 * 0.01 \approx 0 \\ 6 * 10^{-6} * 0.05 * 0.01 = 3 * 10^{-9} \end{cases}$ $= 3 * 10^{-9} \rightarrow (s_{11}, s_{22}, s_{32})$
الحل الثاني للكلمة الأولى s_{12}=ذَهَبُ	الحل الثاني للكلمة الثانية عَلِي = s_{22}	الحل الثاني للكلمة الثالثة بالكُرَة = s_{32}
أفضل الحلول حتى هذه النقطة L_{12} = P(s_{12}/ بداية الجملة) *P(s_{12}) = 0.05*0.02=0.001	أفضل الحلول حتى هذه النقطة $L_{22} = max\begin{cases} L_{11} * P(s_{22}/s_{11}) * P(s_{22}) \\ L_{12} * P(s_{22}/s_{12}) * P(s_{22}) \end{cases}$ $= max\begin{cases} 0.003 * 0.2 * 0.01 = 6 * 10^{-6} \\ 0.001 * 0.1 * 0.01 = 1 * 10^{-6} \end{cases}$ $= 6 * 10^{-6} \rightarrow (s_{11}, s_{22})$	أفضل الحلول حتى هذه النقطة $L_{32} = max\begin{cases} L_{21} * P(s_{32}/s_{21}) * P(s_{32}) \\ L_{22} * P(s_{32}/s_{22}) * P(s_{32}) \end{cases}$ = $max\begin{cases} 7.5 * 10^{-6} * 0 * 0.05 \approx 0 \\ 6 * 10^{-6} * 0.3 * 0.05 = 90 * 10^{-9} \end{cases}$ $= 90 * 10^{-9} \rightarrow (s_{11}, s_{22}, s_{32})$ أفضل المسارات

الجدول ٣-٢: يوضح الحل الكامل بالرموز لمثال التشكيل «ذهب على بالكرة».

ومن الجدول (٣-٢) يمكن أن نستنبط أفضل الحلول عندما وصلنا للمحطة الأخيرة هو الأكثر احتمالاً، ثم نتتبع أفضل المسارات على الإطلاق. سنجد أن حل (s_{11},s_{22},s_{32}) هو أفضل المسارات احتمالا، وهذا يعني أن تشكيل الجملة يكون كالآتي:

ذَهَبَ عَلِّي بالكُرَة

وإذا أخذنا جملاً طويلة، ولبعض كلماتها حلول كثيرة تصل للأربعين حلا، سنجد أنَّنا نبلغ هدفنا في أقل من ثانية، لأننا كلما تقدمنا في الحل نحسب أفضل المسارات من أول كلمة إلى النقطة التي نحن عليها دون عناء البدء دائما من الأول، بل نبدأ من الكلمات التي تسبقنا فقط.

وفي الحقيقة هناك حلول أفضل حتى من فيتربي، من أبرزها (A*Search) أو «الباحث *» يعتمد على n-gram أعلى من bi-gram، ويعطي نتائج أفضل بكثير. كما أن هناك أيضا مصنفات أخرى مثل الشبكات العصبية وآلات الدَّعم المَوَجَّهة (Support Vector Machine) وهي مصنفات حديثة نسبيًّا وذات مقدرة هائلة، لولا ما تحتاجه من إمكانيات عالية، سواء في الذاكرة أو القدرة الحسابية. انظر الملحق - ٣.

٥- مسائل أخرى متشابهة

هناك مسائل لغوية أخرى كثيرة لها نفس الشَّكل الرياضيّ الذي نواجهه عندما تصدينا لحل مشكلة التشكيل الآلي. إن لدينا مستويات مختلفة لمشكلات اللغات الحية عمومًا واللغة العربية خصوصًا، بدءًا من التشكيل الآلي أو الحسم الفونولوجي إلى الحسم الدِّلالي على مستوى الجملة. كلها تشترك في أن الحل يكمن في السياق. ومن هذه المسائل:

٥,١- التشكيل الآلي لبنية الكلمة العربية (المشكلة سالفة الذكر).
(body-word) Diacritization Automatic)) .

٥,٢- التشكيل الإعرابي الآلي للكلمة العربية Automatic Diacritization.

٥,٣- التحليل الصرفي للكلمة العربية (Morphological analysis).

٥,٤- التحليل التركيبي أو النحوي للجملة العربية (Automatic Parsing).

٥,٥- التعرف على «أسماء الكَائنات» من السياق (مثل: أسماء الأعلام، والأماكن، وأسماء المؤسسات، والأحداث، ... إلخ) (Named Entity Recognition).

٥,٦- فكّ الالتباس الدِّلاليّ للكلمات (Word Sense Disambiguation).

وهكذا نجد العديد والعديد من التقنيات التي لها نفس روح المشكلة الرياضية ونفس روح الحل الرياضي مع اختلافات بسيطة من مشكلة إلى أخرى.

٦- أفضل ما سُجِّلَ من نتائج

وهنا لا بد أن نسجل ملحوظة مهمة، وهي أهمية أن تكون هناك وسيلة لقياس النتائج لأنظمة مختلفة بنفس الطريقة وباستخدام نفس البيانات حتى نوحد مسطرة القياس إن جاز التعبير.

وفي مجال التشكيل الآلي، تمت تجارب في كلٍّ من IBM، وجامعة كولومبيا، وشركة RDI، وكانت النتائج المعلنة كما هو موضح في الجدول الآتي:

النموذج	كل تشكيل الكلمة		تشكيل الكلمة عدا الحرف الأخير	
	نسبة الخطأ			
	في الكلمة	على مستوى الحروف	في الكلمة	على مستوى الحروف
نموذج مقدم من د/ عماد زيتوني مع فريق عمل في IBM سنة ٢٠٠٦	٪١٨	٪٥,٥	٪٧,٩	٪٢,٥
نموذج مقدم من د/ نزار حبش مع فريق عمل من جامعة كولومبيا سنة ٢٠٠٧	٪١٤,٩	٪٤,٨	٪٥,٥	٪٢,٢
نموذج مقدم من د/ محسن رشوان مع فريق عمل من شركة RDI سنة ٢٠٠٩	٪١٢,٥	٪٣,٨	٪٣,١	٪١,٢

الجدول ٣-٣: نتائج التَّشكيل الآليّ (IBM ، جامعة كولومبيا ، RDI).

ربما تكون هناك أنظمة أخرى أفضل، ولكن لم يعلن عنها ولم تحّكم على نفس قاعدة البيانات.

وتجدر الإشارة إلى أنَّ قاعدة البيانات قد احتوت على ٢٨٨ ألف كلمة للتدريب و٥٢ ألف كلمة للاختبار. وباختصار قاعدة البيانات وجدنا أن كلمات الاختبار وكلمات التدريب من نفس الطبيعة ومتقاربة جدا. ولما حاولنا تجربة بعض هذه الأنظمة على نصوص أخرى وجدنا أن نسبة الخطأ ربما تصل إلى ٢٥٪ أو حتى ٣٠٪ في بعض الأحيان.

ولكن لحسن الحظ، فإن تشكيل بنية الكلمة أهم بكثير من تشكيلها الإعرابي، ونسبة الخطأ فيه أقلّ كثيرًا. فأذُن الإنسان العربيّ - غير المتخصص في اللغة الآن -لم تعد (وللأسف) حساسة لأخطاء التشكيل الإعرابي قدرَ حساسيتها لتشكيل بنية الكلمة. فالخطأ في تشكيل بنية الكلمة يغير الكلمة ومعناها تغيرا كبيرا على الأذن، فخذ هذا المثال: الفرق بين (كَتَبَ وكُتُب) فرق كبير. فالكلمة الأولى فعل ماض والثانية اسم، وإذا استخدمنا كلمة منهما مكان أخرى، فسيضطرب المعنى في أذن السامع كثيرا، حتى لو تمكَّنَ من استنباطه لاحقًا.

وجدير بالذكر أنَّ هناك توجُّهات حديثة للاستفادة من التَّطوُّر الهائل والحادث في مجال تعلُّم الآلة العميق لحل مُشكلة التَّشكيل الآليّ وما شابهَ من مُشكلات. ويُحاولُ البعضُ مُعالجة مشكلة التشكيل عبرَ التَّرجمة الآليَّة؛ وفي هذه الحالة يُعتبر النَّصُّ الخام والنَّصُّ بعدَ التَّشكيل كلغتين، والمطلوب إجراء ترجمة آلية من النص الخام إلى النص المشكَّل. وفي كل الأحوال تحتاج الوسائل الحديثة والعميقة لتعليم الآلة كميات ضخمة من النصوص المشكَّلة حتى تتمكن من التعلم وإعطاء نتائج جيدة.

٧- طبيعة الموارِد اللُّغَويَّة الَّتي نحتاجها

في الحلول المستخدمة في هذا الفصل نحتاج إلى موارِد لغوية عُولِجَت معالجة خاصة. ففي حالة التشكيل الآلي نحتاج إلى مدونة مشكلة بالكامل، أي: كل حرف فيها مشكل؛ وليس على تشكيل جزئيّ لبعض الحروف التي تفك الالتباس بالنسبة للقارئ العربي. ولكي نصل إلى دقة مناسبة نحتاج لمدونة مشكلة كبيرة وتغطى المجالات المطلوب التشكيل لنصوصها. وفي المجال الواحد قد نحتاج لمدونة بالملايين من الكلمات حتى نتمكن من مقابلة معظم الكلمات المستخدمة في المجال، إذ أن أكبر سبب للأخطاء في

الطرق المعتمدة على تعلم الآلة هو عدم رؤيته للكلمة من قبل بالكلية أو مرت عليه في المدونة في سياقات مختلفة تمامًا.

المدونات الكبيرة جدًّا (ربما بعشرات الملايين من الكلمات) والتي يمكن أن تخفض نسبة الأخطاء بشكل فعَّال مكلفة جدًّا؛ وفي المقابل لا توجد (حتى الآن) مجموعة متكاملة من القواعد التي يمكن الاعتماد عليها لحل المشكلة. وهناك حلّ وسط، وهو مدونة أصغر نسبيًّا (٣-٥ مليون كلمة مثلاً) مع بعض القواعد المساعدة في تخفيض نسبة الأخطاء. إلا أن القواعد تتطلب الاستعانة بتحليل لغويّ مثل المحلل الصرفيّ، حتى يمكن أن تبنى القواعد على الشواهد اللغوية في هذا التحليل.

ولأيّ نظام للتشكيل فعال لا بد من التعامل مع الظواهر اللغوية كثيرة الورود. ومن هذه الظواهر: ظاهرة الكلمات الأجنبية والمكتوبة باللغة العربية (مثل أسماء الرؤساء بوش وأُوباما .. إلخ)؛ هذه الكلمات قد يصل متوسط ورودها لأكثر من ٥٪ في كثير من النصوص الحديثة. ولأن معظمها أسماء لكائنات (أسماء أشخاص أو مؤسسات أو أماكن.. إلخ) فإنها كثيرة ودائمة التغير. فما ورد منها كثيرًا في المدونة المشكلة يتم حسمه كالكلمات العربية؛ وعدا ذلك فإننا نحتاج لبعض القواعد لتعلم لتشكيلها. وهناك مدرسة عملية تجمع ما ورد في المدونة من كلمات أجنبية قبل وبعد التشكيل وتستخدم واحدة أو أكثر من خوارزمات التعلم الآليّ لتعلم تشكيل ما لم يرد في المدونة.

وفي الختام تجدرُ الإشارةُ إلى أن هناك تقدمًا ملحوظا في مجال استخدام الشبكات العصبية في مُختلَف ميادين حوسبة اللغات الحية، بما في ذلك التشكيل الآلي. ولكن ما زالت نتائج الطرق التقليدية تزاحم نتائج الطرق الحديثة، لأن الطرق الحديثة في حاجة إلى كميات ضخمة من البيانات المشكولة يدويًّا. وهذا ليس سهلا بالنسبة للنصوص المعاصرة. يتضح الفرق حين تُشَكَّل نصوص تراثية، إذ تعطي الشبكات العصبية نتائج أفضل بكثير.

٨- أفكارٌ بحثيَّة لأطرُوحاتٍ علميَّةٍ مُستَقبليَّة

٨,١- تصميم مدونة مشكَّلة صغيرة نسبيًّا؛ ولكن تظل ممثلة جيدة للمجال الذي صُممت له. كيف تختار موضوعاتها حتى تحقق كلماتها أعلى نسبة شمول لكلمات المجال.

٨,٢- استخلاص مجموعة من القواعد الممكن تنفيذها حاسوبيًّا. إن هناك كثيرًا من القواعد لا يمكن تنفيذها حاسوبيًّا. مثال ذلك: الجملة الاسمية تتكون من مبتدأ وخبر، والخبر مع المبتدأ يُتِمَّان المعنى. هذه قاعدة تحتاج لمعرفة معاني الجمل؛ أتَمَّت أم لا؛ وهذا لم نصل له بعد. إنما إذا قلنا: إن الصفة تتبع الموصوف في التعريف والعدد والنوع، فهذه القاعدة يمكن تطبيقها حاسوبيًّا. فبالرجوع إلى محلل صرفي للغة العربية، يمكن معرفة كل المطلوب، وبالتالي يمكن حسم الصفة بالقواعد.

٨,٣- دراسة مدونة مشكلة آليًّا وتحليل الأخطاء الناجمة عن المشكل الآلي، ثُمَّ وضع القواعد التي تقلل هذه الأخطاء. هذا البحث يمكن أن يؤدي إلى نتائج أفضل عمليًّا من كل الحلول المتاحة. ومن الخبرة في هذا المجال أن عددًا قليلاً من القواعد مسئول عن نسبة كبيرة من الأخطاء. وبديهي أن ذلك يحتاج إلى تحليل وتصنيف للأخطاء قبل وضع القواعد.

٨,٤- عمل مجموعة من القواعد التي يمكن استخدامها لتشكيل الكلمات الأجنبية الواردة في النصوص الحديثة. هذا البحث يمكن أن يساعد على حسم عدد لا بأس به من الكلمات الحديثة. إن هذه دراسة تتداخل فيها الصوتيات العربية واللاتينية.

ببليوجرافيا مرجعيَّة

1. Abandah, G.A.; Graves, A.; Al-Shagoor, B.; Arabiyat, A.; Jamour, F.; Al-Taee, M. (2015). "Automatic diacritization of Arabic text using recurrent neural networks". International Journal on Document Analysis and Recognition (IJDAR), Vol. 18, pp. 183-197.
2. Attia, M. (2000). "A Large-Scale Computational Processor of the Arabic Morphology, and Applications" M.Sc. thesis, Dept. of Computer Engineering, Faculty of Engineering, Cairo University.
3. Attia, M. (2005). "Theory and implementation of a large-scale Arabic phonetic transcriptor, and applications", Ph.D. dissertation, Dept. of Electron. and Elect. Commun., Faculty of Eng., Cairo Univ., Cairo, Egypt, Sep. 2005.
4. Balivada, L. K.; Raju, K. P. (2012). Optimization Techniques of Viterbi Algorithm: Performance Analysis of Different Algorithms. LAP.
5. Darwish, K.; Mubarak, H. (2016). "Farasa: A new fast and accurate Arabic word segmenter", In Proceedings of the Tenth International Conference on Language Resources and Evaluation 2016, Paris, France, may. European Language Resources Association (ELRA).
6. Deng, L.; Liu, Y. (2018). Deep Learning in Natural Language Processing. Springer.
7. Emam, O.; Fisher, V. "A Hierarchical Approach for the Statistical Vowelization of Arabic Text", U.S. patent application US2005/0192809 A1, 2004, IBM patent filed, DE9-2004-0006.
8. Habash, N.; Rambow, O. (2007). "Arabic diacritization through full morphological tagging", in Proc. 8th Meeting North Amer. Chap. Assoc. Computat. Linguist. (ACL); Human Lang. Technol. Conf. (HLT-NAACL).
9. Habash, N.; Shahrour, A.; Al-Khalil, M. (2016). "Exploiting Arabic Diacritization for High Quality Automatic Annotation". Language Resources and Evaluation Conference, Portoroz, Slovenia.

10. Information Resources Management Association (2011): Machine Learning: Concepts, Methodologies, Tools and Applications. IGI Global Snippet.

11. Kaleli, C.; Polat, H. (2010). NAÏVE BAYESIAN CLASSIFIER-BASED PRIVATE RECOMMENDATIONS: PRIVACY-PRESERVING NAÏVE BAYESIAN CLASSIFIER-BASED COLLABORATIVE FILTERING. LAP.

12. Karwowski, W. (2019). Intelligent Human Systems Integration 2019. Springer.

13. Kulkarni, A.; Shivananda, A. (2019). Natural Language Processing Recipes: Unlocking Text Data with Machine Learning and Deep Learning using Python. Apress.

14. Meghanathan, N.; Kaushik, B. K.; Nagamalai, D. (2011). Advances in Computer Science and Information Technology: First International Conference on Computer Science and Information Technology, CCSIT 2011, Bangalore, India, January 2-4, 2011. Proceedings.

15. Metwally, A.; Rashwan, M.; Atiya, A. (2016): A multi-layered approach for Arabic text diacritization. IEEE International Confer-ence on Cloud Computing and Big Data Analysis (ICCCBDA).

16. Neamat El, G.; Yee, S. (2018). Computational Linguistics, Speech and Image Processing for Arabic Language. World Scientific.

17. Pasha A., Al-Badrashiny M., Diab M., El Kholy A., Eskander R., Nizar Habash, Pooleery M., Rambow O., and Roth M., "Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic". In LREC-2014, Reykjavik, Iceland.

18. Rashwan M., Al Sallab A., Raafat M. and Rafea A., "Deep learning framework with confused sub-set resolution architecture for automatic Arabic diacritization". In IEEE Transactions on Audio 2015, Speech and Language Processing, pages 505-516.

19. Rashwan, M. A. A.; Al-Badrashiny, M. A. S.; Attia, M.; Abdou, S. M.; Rafea, A.; "A Stochastic Arabic Diacritizer Based on a Hybrid of Factorized and Unfactorized Textual Features"; IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 19, NO. 1, JANUARY 2011.

20. Sammut, C.; Webb, G. I. (2011). Encyclopedia of Machine Learning. Springer.

21. Shaalan, k.; Hassanien, A.; Tolba, F. (2017). Intelligent Natural Language Processing: Trends and Applications. Springer.

22. Sharp, B.; Sedes, F.; Lubaszewski, W. (2017). Cognitive Approach to Natural Language Processing. Elsevier.

23. Shen, G.; Huang, X. (2011). Advanced Research on Computer Science and Information Engineering: International Conference, CSIE 2011, Zhengzhou, China, May 21-22, 2011. Proceedings.

24. Shi, Z. (2011). Advanced Artificial Intelligence. World Scientific.

25. Srinivasa-Desikan, V. (2018). Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras. Packt Publishing.

26. Vasquez, D.; Gruhn, R.; Minker, W. (2013). Hierarchical Neural Network Structures for Phoneme Recognition. Springer.

27. Zaghouani, W.; Bouamor, H.; Hawwari, A.; Diab, M.; Obeid, O.; Ghoneim, M.; Alqahtani, S.; Oflazer, K. (2016): Guidelines and Framework for a Large Scale Arabic Diacritized Corpus. Proceedings of the International Conference on Language Resources and Evaluation (LREC'2016).

28. Zitouni, I.; Sorensen, J. S.; Sarikaya, R. "Maximum entropy based restoration of Arabic diacritics", in Proc. 21st Int. Conf. Comput. Linguist. and 44th Annual Meeting Assoc. for Comput. Linguist. (ACL); Workshop Comput. Approaches to Semitic Lang., Sydney, Australia, Jul. 2006 [Online]. Available: http://www.ACLweb.org/anthology/P/P06/P06-1073.

الفصل الرابع
التَّنقيب في النُّصُوص

د. علي علي فهمي

المبحث الأوَّل: التَّجميع والتَّصنيف.
المبحث الثَّانِي: تلخيص النُّصُوص.
المبحث الثَّالث: استنباط اتِّجاهات الرَّأي العامّ.



الشَّكل ٤-١: التنقيب عن البيانات يماثل التنقيب عن المعادن الثمينة.

مصطلح التَّنقيب في النُّصُوص - والبيانات بصفة عامة - مأخوذ من مصطلح التَّنقيب في المعادن الثمينة وسط تلال من الأشياء. فالتَّنقيب في النُّصُوص كما عرَّفته «مارتي هيرست» - من جامعة بيركلي بكاليفورنيا في ٢٠٠٣م - هو استخدام الحاسوب في اكتشاف معلومات غير معروفة مسبقا من مصادر متنوعة من النصوص، ولكنها موجودة في ثنايا هذه النصوص. ومن هذه المعلومات - مثلا - التعرف على أسباب بعض الأمراض النادرة من خلال فحص وثائق العلوم الحيوية المختلفة، واكتشاف البروتينات التي تتفاعل مع غيرها من البروتينات الأخرى، وهي خاصية مُهِمَّة جدًّا، تُؤخذ في الاعتبار عند تصنيع الدواء وعند وصف العلاج.

ويعتبر التَّنقيب في النصوص أحد علوم الحاسب الحديثة، وترجع نشأته إلى منتصف السبعينيَّات عندما اقترح «جيرارد سالتون» - من جامعة كورنيل - تمثيل النصوص المكتوبة باللغات الطبيعية بواسطة متجهات رقمية والتعامل معها بالأساليب الرياضية المستخدمة في التعامل مع المصفوفات العددية والأساليب المستخدمة في التعامل مع قواعد البيانات النمطية. وقد مكَّنَ التقدم التكنولوجي هذا المجال من المضي قدما خلال العقد الماضي بصورة ملموسة.

وكانت أبحاث العالم الأمريكي «دون سوانسن» - من جامعة شيكاغو - علامة فارقة في مولد علم التَّنقيب في النُّصُوص.

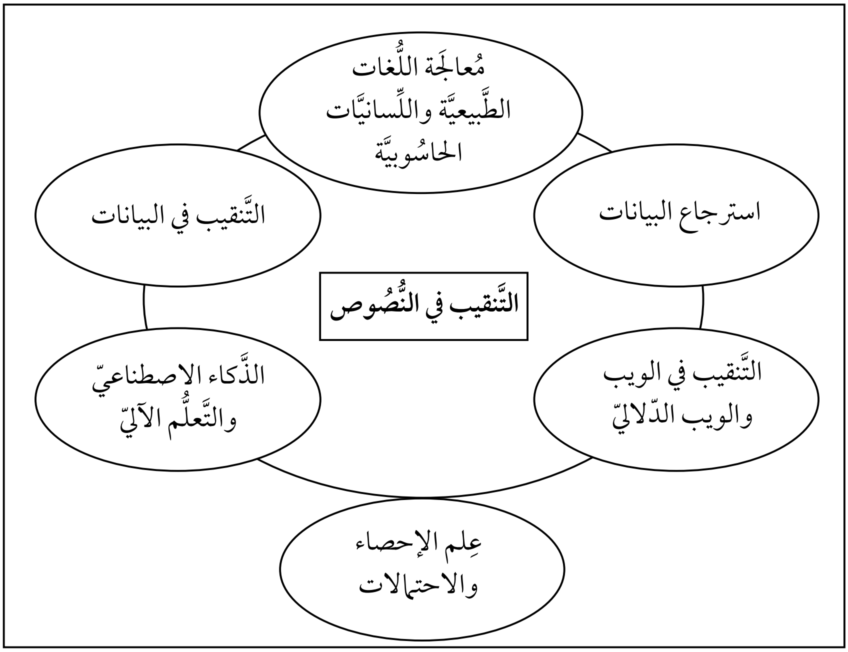
فقد لاحظ «سوانسن» ضعف التواصل العلمي بين المتخصصين في حالة اختلاف مجالاتهم العلمية الدقيقة وعدم المعرفة بما يدور في المجالات الأخرى، وبالتالي عدم الاستفادة بها. ولذا قام - بالاشتراك مع زميله نيل سماهيسر - ببناء نظام لاكتشاف سلاسل من النتائج والآثار السببية من خلال فحص الدوريات العلمية في التخصصات المختلفة. ونتيجة لذلك فقد اكتشف «دون سوانسن» في ١٩٩٧ أن نقص الماغنيسيوم في جسم الإنسان هو من الأسباب الرئيسية للصداع الذي يصيبنا. هذه المعلومة لا توجد في

أي كتاب أو دورية علمية ولكن تم استنباطها بواسطة نظام التَّنقيب بالبحث التتابعيّ في دوريات وقواعد البيانات الطبية وفي دوريات ومنشورات المعلوماتية الحيوية عن أسباب الأمراض. في حالة أمراض الصداع تم التوصل إلى ثمان سلاسل سببية من خلالها تم الربط بين الصداع وبين نقص الماغنيسيوم.

وجدير بالذكر أن التقنيات والأساليب المستخدمة في التَّنقيب في النُّصُوص تستخدم أيضا في مجالات أخرى عديدة، مثل مجالات التَّنقيب في البيانات، ومجالات التعرف على الصور، ومجالات التعرف على الكلام، ومجالات التعرف على الكتابة. وبالطبع يفيد كلٌّ منها الآخر.

ويختِلفُ التَّنقيب في النُّصُوص عن البحث في النصوص أو البحث في صفحات الإنترنت بواسطة برمجيات ومحركات البحث العالمية الشهيرة، مثل: جوجل (Google) وياهو (Yahoo) وبينج (Bing). فعند استخدام محركات البحث يبحث المستخدم عن شيء معروف قد تم إعداده مسبقًا بواسطة آخرين، كأن يبحث عن عنوان شركة تنتج منتجًا بمواصفات معينة، أو يبحث عن أول أمين عام للأمم المتحدة، أو عن الدول التي انضمَّت إليها حديثا خلال آخر ثلاث سنوات، وهكذا. وبالطبع فإن محركات البحث تغني المستخدم عن البحث في مئات بل آلاف الوثائق غير ذات العلاقة.

التَّنقيب في النُّصُوص هو مجال متعدد التخصصات، يعتمد على علوم استرجاع المعلومات والبيانات، وعلوم التَّنقيب في البيانات العددية، وعلوم الذكاء الاصطناعي والتعلم الآلي، وعلوم الإحصاء والاحتمالات، وعلوم معالجة اللغات الطبيعية واللُّغويات الحاسوبية، وذلك على النَّحو الموَضَّح بالشكل التالي:



الشَّكل ٤-٢: التخصصات المشاركة في مجال التنقيب في النصوص.

تطور مفهوم التَّنقيب في النُّصُوص في الآونة الأخيرة ليشمل تطبيقات أخرى غير نمطية لا تشملها محركات البحث، مثل: التصنيف الآلي لآلاف - بل ملايين - الوثائق إلى وثائق سياسية، صحية، اجتماعية، رياضية، فنية، وغيرها، بدون الحاجة إلى الاستعانة بالمختصين؛ ومثل: تجميع النصوص «في مجموعات مُتشابهة»، وتلخيص الوثائق، والتَّنقيب في الآراء وتحليل المشاعر، والتصحيح والتصويب الآلي للإجابات الإنشائية، واستنباط المفاهيم، والتعلم الآلي للأنطولوجيات، وغيرها من التطبيقات المهِمَّة.

لقد حظيت تقنيات وتطبيقات تصنيف وتجميع الوثائق المتشابهة بالغالبية العظمى من النشر العلمي على مدى الأعوام السابقة. فبالإضافة إلى كونها تطبيقات في حد ذاتها مثل تصنيف البريد الإلكتروني وتصنيف الأخبار، إلا أنها أصبحت مكوِّنًا رئيسيًّا في كثير من تطبيقات التَّنقيب في النُّصُوص كما سيتضح.

سنتناول في هذا الفصل - بشيء من التفصيل - الموضوعات التالية، والَّتي زاد الاهتمام بها في الآونة الأخيرة بصورة كبيرة على المستويين - النظري والتطبيقي:

١- التَّجميع والتَّصنيف.

٢- تلخيص النصوص.

٣- استنباط اتِّجاهات الرَّأي العامّ (التَّنقيب في الآراء).

المبحث الأوَّل
التَّجميع والتَّصنِيف

١ - مُقَدِّمة.

٢- نماذج من التطبيقات العملية للتجميع والتصنيف للنصوص.

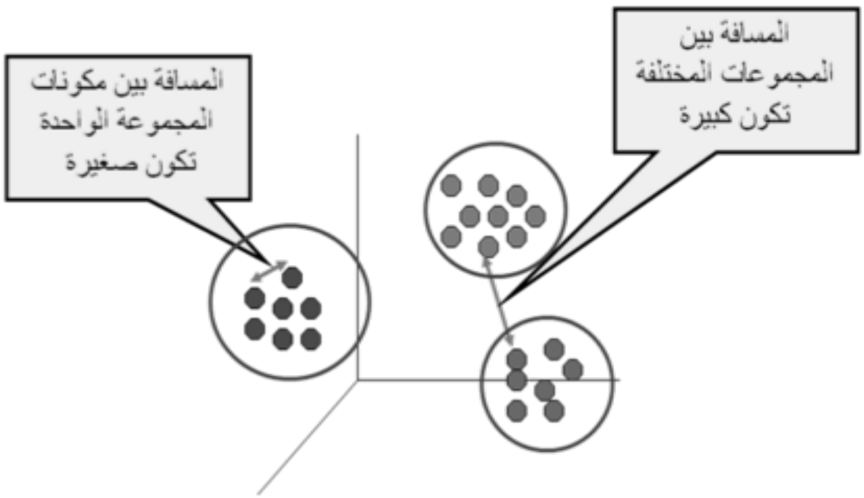
٣- خوارزمات التجميع والتصنيف.

٤- خوارزمات التجميع والتصنيف واللغة العربية.

١ – مُقَدِّمة

١,١ – التجميع (Clustering)

يُمثِّل تجميع البيانات إحدى تقنيات التنقيب في البيانات والنصوص التي تمكن من تقسيم مجموعة كبيرة من البيانات أو فئات من الكائنات إلى عدة مجموعات فرعية ذات خصائص متشابهة أو مغزى مشترك. فعلى سبيل المثال يمكن للفرد العادي تقسيم الجماهير التي تشاهد مباراة لكأس العالم بين ألمانيا والبرازيل إلى ثلاث مجموعات، المجموعة الأولى تشجع الفريق الألماني والمجموعة الثانية تشجع الفريق البرازيلي، أما المجموعة الثالثة فهي من عشاق اللعبة الحلوة ولا تنتمي لأي من الفريقين. وبالمثل أيضا يمكن لنا تقسيم رسائل الماجستير التي تمت إجازتها بقسم الحاسب بجامعة القاهرة إلى عدة مجموعات تعكس المجالات البحثية لهذا القسم العلمي، مع مراعاة أن هذا التقسيم يتم بدون تدخل بشري من المختصين.



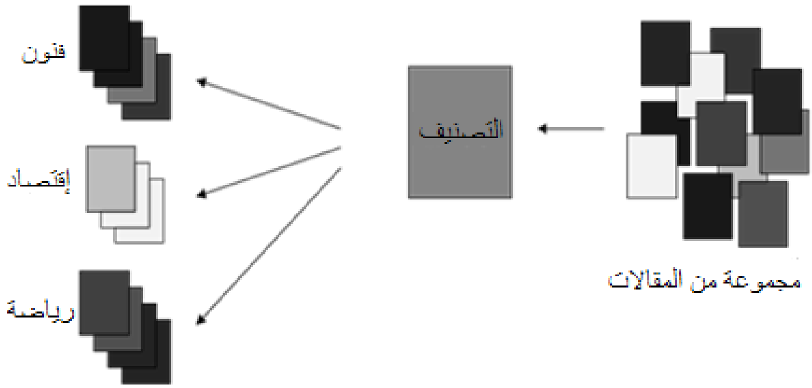
الشَّكل ٤ –٣: تقسيم الكائنات إلى مجموعات متشابهة.

والسؤال المطروح هو: على أي أساس تتم عملية تقسيم البيانات والنصوص إلى مجموعات متشابهة؟ وكيف يمكن للآلة القيام بهذه الأعمال آليا؟

١ , ٢- التصنيف (Categorization)

بفرض وجود مجموعة من الفئات أو الأنواع المعروفة مسبقا فإن عملية تصنيف بيان ما أو كيان ما أو نص ما تتمثل في اختيار الفئة أو النوع الذي ينتمي إليه هذا النص. فمثلا يستطيع القارئ تصنيف المقال الذي يطلع عليه إلى واحد أو أكثر من أنواع المقالات المعروفة: مقالات سياسية، مقالات أدبية، مقالات فنية، وهكذا.

مثال آخر: يستطيع البنك تصنيف عمليات استخدام كروت الائتمان إلى عمليات سليمة وعلميات تمت من خلال النصب والاحتيال.



الشّكل ٤-٤: تصنيف المقالات الإخبارية طبقا لموضوعاتها

يلاحظ هنا أن عملية التصنيف تختلف عن عملية التجميع من حيث فرضية وجود أنواع معروفة مُسبقًا، أما التجميع فلا يفترض ذلك.

حقيقة الأمر أننا نمارس عمليات التجميع والتصنيف في جميع الأوقات في حياتنا اليومية، فعندما نستمع إلى صوت خارج من المذياع فإننا نصنفه إلى صوت ترتيل القرآن أو صوت تحليل إخباري أو صوت موسيقى مثلا. وعندما تقابل شخصًا لأول مرة فإنك حتما ستقوم بتصنيفه من حيث المستوى الاجتماعي أو المستوى الثقافي أو المستوى الجمالي أو المستوى العلمي أو إلى غيره من المستويات، دون الشعور أو تعمد ذلك. وبمقابلة أعداد كثيرة من الأشخاص قد تتكون لديك الرغبة في تقسيمهم إلى مجموعات تختلف عن التصنيف الشائع بين الأفراد.

انظر إلى الصورة التالية، والتي بها مجموعة من الأشخاص. كيف يتم تقسيمهم إلى مجموعات؟ وما عدد هذه المجموعات؟ وعلى أي أساس تم التجميع؟



وإلى أي مجموعة ينتمي الشخص التالي:



والسؤال المطروح دائما هو: على أي أساس تتم عملية تحديد انتماء شيء أو بيان أو نص ما إلى واحدة أو أكثر من الأنواع المعروفة مسبقا؟ وكيف يمكن للآلة القيام بهذه العملية آليا؟

وسؤال آخر مطروح هو: ما هي التطبيقات العملية لهذه التقنيات؟

قبل الإجابة على هذه الأسئلة يجب علينا تمييز أصناف ونوعية البيانات التي تتم عليها عمليات التجميع والتصنيف إلى الأنواع التالية (ويطلق عليها اسم الوسائط المتعددة Multimedia):

- الأفلام المرئِيَّة.
- الصور والرسوم المتحركة.

- الكلام المنطوق.
- الكلام المكتوب (النصوص المكتوبة).
- قواعد البيانات والجداول الرقمية.

ومع الاختلاف الواضح بين هذه البيانات إلا أنَّ التقنيات المستخدمة في تجميع وتصنيف هذه النوعيات المختلفة من البيانات تتشابه إلى حد كبير، بل تتطابق في كثير من الأحيان؛ إلا أن الاختلاف الجوهري بينها يكون في طبيعة السِّمات (Features) التي يتم التجميع والتصنيف بناءً عليها.

سوف نهتم في هذا الفصل بالتجميع والتصنيف للكلام المكتوب (النصوص الكتابية)؛ وهو الأساس النَّظريّ والعَمَليّ لجميع التَّطبيقات المنبَثِقة عن التنقيب في النصوص (Text Mining).

٢- نماذج من التطبيقات العملية للتجميع والتصنيف للنصوص

أصبحت منتجات التنقيب في النصوص متاحة الآن للاستخدامات العملية وليست مقصورة على مستوى المراكز البحثية، ولا يكاد يخلو تطبيق الآن من استخدام تقنيات التجميع والتصنيف؛ ونذكر منها:

٢,١- تطبيقات في مجال الأمن

مثل التصنيف الآلي لدرجات السرية للوثائق (سري، سري جدا، سري للغاية، محظور، بدون).

٢,٢- تطبيقات في مجال الطب الحيويّ

تستخدم تقنيات التجميع والتصنيف في بناء آلات البحث الدلالية، مثل: (GoPubMed and GoPubmed)، والتي تستخدم في البحث عن الجينات وعرض النتائج في صورة شجرية.

٢,٣- تطبيقات التنقيب في الشبكات الاجتماعية

ويستفاد من هذه التطبيقات في شركات الدعاية والإعلان الانتقائي، كما تستفيد منها المؤسسات الأمنية في تتبع الأشخاص من خلال العلاقات الاجتماعية الخاصة بأقرانهم

على الشبكات الاجتماعية. وتطبيقات التنقيب في الآراء على الشبكات الاجتماعية تعتبر من الموضوعات الحديثة الجاذبة لكثير من الباحثين لما لها من تأثير مباشر سواء على المستوى التجاري أو التكنولوجي.

٢,٤- تطبيقات في مجال التسويق التجاري

تستخدم هذه التطبيقات في تحليل العلاقة بين الشركات والعملاء وبناء أنظمة تنبؤ مبنية على ذلك. وهناك آلات بحث دلالية غير التي سبق ذكرها، مثل آلات: (Find TheBest, Hunch, Pikimal)، وتستخدم في دعم عملية الاختيار التي يقوم بها المستخدم عند قيامه بالشراء من خلال الإنترنت.

٢,٥- تطبيقات في المجال الأكاديمي

تمثل تقنيات التجميع والتصنيف ضرورة مهمة لمؤسسات النشر الأكاديمية، والتي لديها مئات الآلاف أو أكثر من الكتب والمجلات والمنشورات العلمية التي تحتاج إلى فهرسة لاسترجاعها، مع الأخذ في الاعتبار نشأة العلوم والمجالات العلمية الجديدة مما يتطلب تحديث الفهارس أولاً بأول.

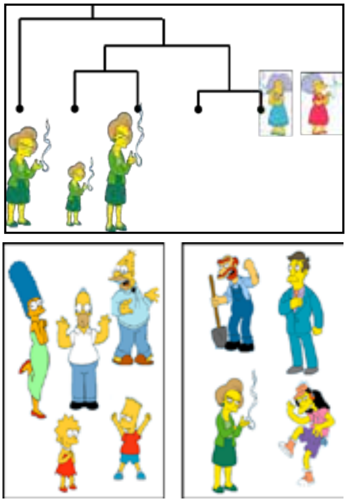
لذلك اهتمت بهذا الموضوع مراكز قومية عديدة، مثل المركز القومي للتنقيب في النصوص بالمملكة المتحدة (NaCTeM) ومؤسسات نشر عالمية، مثل مؤسسة نشر مجلة «الطبيعة» الشهيرة (Nature Journal) والمعاهد الطبية الوطنية للصحة بالولايات المتحدة وكثير من الجامعات الَّتي قامت بإطلاق عدة مبادرات مهمة في مجال توصيف الوثائق وفي مجال بناء واجهات الاستخدام والبحث، مثل مبادرة بناء واجهات مفتوحة المصدر مبنية على التنقيب في النصوص (OTMI – Open Text Mining Interface)، ومبادرة تعريف نوع المستند (DTD – Document Type Definition) والتي من شأنها توفير إشارات دلالية للآلة في الإجابة على أسئلة محددة وردت في نص الوثيقة.

٢,٦- تطبيقات التصنيف الآلي للبريد الإلكتروني

وتفيد هذه التطبيقات في التعرف وحجب مئات الرسائل الإلكترونية التي تصلنا يوميا من مصادر مجهولة أو تحتوي على موضوعات غير مرغوبة فيها. وعموما فإن هذه التطبيقات تستخدم في تصنيف الرسائل الإلكترونية إلى بريد مهم جدا، وبريد مهم، وبريد عادي، وبريد غير مرغوب فيه، وبريد دعائي، إلخ.

٢,٧- تطبيقات تجميع نتائج البحث في الإنترنت وتصنيفها

على سبيل المثال، عندما نبحث عن كلمة «خلية cell» تقوم آلات البحث، مثل آلة البحث (Vivisimo) وآلة البحث (Northern Light)، بتجميع الآلاف من نتائج البحث وتقسيمها آليا إلى مجموعات متشابهة في المجال مثل: الخلايا البيولوجية، والخلايا الشمسية، والخلايا الإرهابية؛ مما يساعد الباحث في الوصول إلى ما يريده من معلومات.



٣- خوارزمات التجميع والتصنيف

توجد خوارزمات عديدة لعمليات التجميع والتصنيف وتختلف من حيث نظرية عملها ودقة النتائج التي تصل إليها. منها ما يحتاج إلى تدريب وتعلم مسبق، ومنها ما لا يحتاج. ونذكر فيما يلي أهم هذه الخوارزمات وأكثرها انتشارًا. يمكن وصف خصائص مختلفة من خوارزمات المجموعات على النحو التالي:

٣,١ - خوارزمات التجميع (Clustering)

- **التجميع الهرمي (Hierarchical Clustering)**

توصف كل مجموعة بأكبر مسافة مسموح بها بين كل عنصر وآخر من عناصرها. يتم حساب المجموعات بأحد أسلوبين:

الأسلوب الأول هو الأسلوب التجميعي حيث يتم اعتبار كل عنصر مجموعة قائمة بذاتها ثم يتم دمج هذه المجموعات الصغيرة إلى مجموعات أكبر. وتتوق عملية الدمج إذا تم الإخلال بشرط أكبر مسافة مسموح بها. والأسلوب الآخَر هو الأسلوب التقسيمي حيث يتم وضع جميع العناصر في مجموعة واحدة ثم يتم تقسيم هذه المجموعة إلى مجموعات فرعية في حالة الإخلال بشرط أكبر مسافة مسموح بها.

- **التجميع من خلال حساب مراكز الثقل**

(Clustering means-K) clustering based-Centroid)

تعتمد هذه الطريقة على تحديد عدد المجموعات مسبقا بواسطة المستخدم، ويُختار لكل مجموعة مركز ثقل، ويتم توزيع العناصر على كل مجموعة طبقًا لبُعد العنصر عن مركز الثقل، ثم يعاد حساب مراكز الثقل مرة ثانية، ويعاد توزيع العناصر مرة ثانية وثالثة وهكذا طالما هناك تغيير في مراكز الثقل. وتنتهي عملية التجميع مع ثبوت مراكز الثقل الجديدة.

- **التجميع من خلال حساب الكثافة (Density-based clustering (DBSCAN)**

ويتم فيه النظر إلى المجموعة على أنها المساحة ذات الكثافة العالية من العناصر، أما العناصر المبعثرة فيتم اعتبارها فواصل أو عبارة عن شوشرة وضوضاء.

٣,٢- خوارزمات التصنيف (Classifications) (من خلال التعلم)

تقوم هذه الخوارزمات ببناء نماذج التصنيف من خلال دراسة مجموعة من الأمثلة لعدة فئات معروفة مسبقا. وبواسطة هذه النماذج يتم تصنيف العناصر الجديدة التي لم تسبق رؤيتها. وأشهر هذه الخوارزمات:

- آلة الدعم الموجهة (Support Vector Machine).
- الشبكات العصبية (Neural Networks) .
- طبقًا لأقرب الجيران (Nearest Neighbors-k).
- طريقة بايز المبَسَّطة (Naive Bayes).
- شجرة القرار (Decision Tree).
- شبكات بايز (Bayesian Networks).

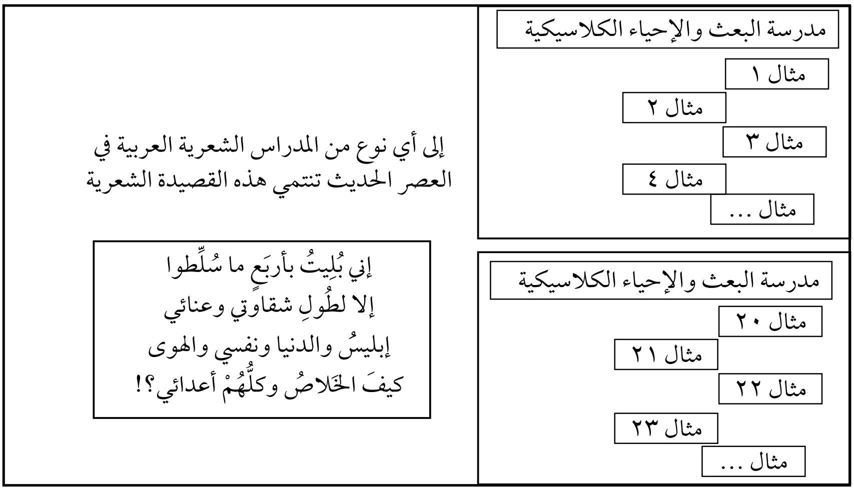
٤- خوارزمات التجميع والتصنيف واللغة العربية

تعتمد جودة التجميع والتصنيف على اختيار واستخلاص ملامح/ سمات العناصر التي تتم تغذيتها للخوارزمات المذكورة سابقا.

يوجد اتجاهان لأخذ خصائص اللغة العربية في الاعتبار عند بناء التطبيقات الخاصة بالتجميع والتصنيف وبناء تطبيقات التنقيب في النصوص بصفة عامة.

الاتجاه الأول هو استغلال الخوارزمات التي تم تطويرها للعمل في بيئة اللغة الإنجليزية بدون تغيير، ويتم التركيز على اختيار السِّمات التي تأخذ في الاعتبار خصائص الصرف والنحو العربي.

والاتجاه الآخَر (وهو التوجه الحاليّ)؛ يتمثل في الخوارزمات التي تم تطويرها للعمل في بيئة اللغة الإنجليزية من خلال تعديل المعادلات المستخدمة داخل هذه الخوارزمات أثناء حساب المسافات بين العناصر لتأخذ في الاعتبار خصائص اللغة العربية. المثال التالي يوضح أوجه الاختلاف في التعامل بين النص الإنجليزي والنص العربي عند تطبيق خوارزمات التصنيف؛ والمطلوب بناء برنامج قادر على: تحديد إلى أي المدارس الشعرية تنتمي قصيدة شعرية معينة، علما بأنه لدينا أمثلة عديدة من القصائد التي تنتمي لكل مدرسة شعرية.



الشَّكل ٤-٥: مثال يوضح أهمية أخذ خصائص اللغة العربية في الإعتبار عند بناء تطبيقات التنقيب في النصوص العربية

في حالة القصيدة العربية:

أحد السِّمات التي يتم تغذية خوارزم التصنيف بها يتمثَّل في الأوزان التي جاءت عليها الأفعال داخل القصيدة، وهو ملمح يختص باللغة العربية فقط.

في حالة القصيدة الإنجليزية، يمكن الأخذ في الاعتبار ملامح أخرى تختص باللغة الإنجليزية، مثل مدى استخدام الصيغ المختصرة (Yr instead of Your) كما هو مستخدم في قصائد «الجبل الأسود Black Mountain».

المبحث الثَّاني

تلخيص النُّصُوص

١- أنواع التلخيص الآلي.

٢- قياس جودة التلخيص الآلي.

٣- أساليب التلخيص الآلي.

٤- نماذج من أنظمة التلخيص الآلي.

٥- الخُلاصة.

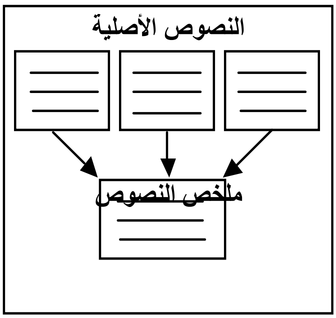
مع تزايد إنتاج الوثائق الإلكترونية بصورة تصاعدية وإتاحتها على شبكة الإنترنت كل يوم، كان من الضروري العمل على إيجاد أنظمة تساعد في تلخيص واستخلاص محتوى هذه الوثائق والاستفادة من المعلومات الموجودة بها. ومن شأن هذه الأنظمة توليد ملخص لمستند أو مجموعة من المستندات، أو تلخيص جملة طويلة، مع حذف المعلومات الزائدة والمكررة والحدّ من التفاصيل.



عندما نبحث عن معلومة أو بيان على الإنترنت باستخدام محركات البحث، ففي معظم الأحوال يقوم محرك البحث بالإفادة بأن هناك مئات الآلاف من الوثائق التي تحتوي على الكلمات الحاكمة التي تعبر عن المعلومات المطلوبة، ويقوم باختيار وإظهار عدد قليل منها، في حدود العشرات، حيث يتم اختيارها وترتيبها بناء على خصائص لا ترتبط بالمعنى أو المحتوى النصي للوثيقة. وبالطبع لا يستطيع القارئ الاطلاع على كل أو حتى جزء صغير منها. تساعد أنظمة تلخيص النصوص في تلخيص هذه الوثائق وعرض هذه الملخصات حيث يستطيع القارئ أن يحدد الوثيقة أو الوثائق التي يتفحصها تفصيليا. كذلك يمكننا أن نتعرف على ملخص الموضوعات التي تداولناها مع أحد الأشخاص أو إحدى الجهات من خلال البريد الإلكتروني خلال العام الماضي مثلا.

مثال آخر يتمثل في عرض ملخصات الأخبار على أجهزة التليفونات المحمولة حيث شاشتها الصغيرة تجعل من غير الملائم عرض الخبر بالكامل.

ويعرف ملخص النص بأنه الموجز الذي يتم إنتاجه من واحد أو أكثر من النصوص، ويحتوي على نسبة كبيرة من المعلومات الموجودة في النص الأصلي ولا يتعدى حجمه نصف النص الأصليّ.



كيف تتمّ عملية تلخيص النصوص؟

قبل الدخول في شرح تقنيات التلخيص الآلي للنصوص.. هناك بعض الأسئلة التي تطرح نفسها مثل:

١- أي الأنواع من التلخيص يحتاجها المستخدم.

٢- كيف نقيس جودة التلخيص الآلي.

١- أنواع التلخيص الآلي

يمكن النظر إلى نوعية التلخيص من عدة زوايا مختلفة. فمن حيث الغرض منه: هناك تلخيص معبر عن النص، وهناك تلخيص كمؤشر عن نوعية النص. النوع الأول هو الأكثر دقة، أما النوع الآخَر فيستخدم للتصنيف الآلي للنص.

ومن حيث طبيعة الملخص الناتج؛ هل هو أجزاء مستقطعة من النص الأصلي أم هو إعادة صياغة للنص الأصلي محتفظا بمعناه ولكن في سطور أقل بنسبة ٩٠٪ مثلا.

ومن حيث رغبة المتلقي، هل الملخص يعكس رؤية النص الأصلي أم يعكس ما يهتم به القارئ، كما تعتمد طبيعة التلخيص الآلي على الشخص المتلقي للمعلومة، فمثلا تلخيص نص سياسي لقارئ متخصص يختلف عنه لقارئ عادي، وتلخيص قصة لشخص صغير السن يختلف عنه لشخص كبير السن.

ومن حيث مصدر النصوص المطلوب تلخيصها ومن حيث اللغة المكتوبة بها، فهل المصدر وثيقة واحدة أم عدة وثائق، وهل المصادر مكتوبة جميعا بنفس اللغة (العربية مثلا) أم بعدة لغات مختلفة.

طبقا لنوع التلخيص، هناك أسلُوبان رئيسيان للتلخيص الآلي: أسلُوب استخلاص عدد محدود من الجمل من النص أو النصوص الَّتي يتم اختيارها طبقا لمعايير معينة، وأسلُوب إعادة صياغة النص بجمل في الغالب تكون جديدة ومختصرة يتم الوصول إليها من خلال تفهم النص أو النصوص الأصلية. ونظرا لصعوبة عمليات فهم النصوص ونجاحها في مجالات تخصصية محدودة، فإن أسلُوب استخلاص عدد محدود من الجمل من النص هو الشائع حاليا في مجال تلخيص النصوص.

٢- قياس جودة التلخيص الآلي

تقاس جودة التلخيص من خلال عنصريين أساسيين:

٢,١- **نسبة ضغط النص**: ويعبر عنها بطول الملخص مقارنة بالنص الأصلي، ويقصد بالطول هنا عدد كلمات أو عدد الجمل أو عدد الفقرات الموجودة بالملخص، وهذه النسبة يسهُل حسابها من خلال معادلات بسيطة. ونسبة

ضغط النص عادة يتم تحديدها أو اختيارها مسبقا بواسطة المستخدم قبل تنفيذ عمليات التلخيص.

٢,٢- نسبة الاحتفاظ بالمعلومة.

ولكن:

كيف نحدد ما إذا كان الملخص قد احتفظ بكامل المعلومة الأساسية الموجودة في النص الأصلي أم لا؟ وما هي نسبة الاحتفاظ؟

توجد عدة طرق للتعامل مع هذه المعضلة الكبيرة:

- الطريقة الأولى: تعتمد على فحص ناتج التلخيص والحكم على جودته بواسطة المختصين، وبالطبع نتيجة الحكم تختلف من شخص إلى آخر.
- الطريقة الثانية: تعتمد على حساب عدد مقاطع الكلمات المشتركة بطول معين وفقًا للنَّحو العَدَدِيّ (N-gram) بين ناتج التلخيص الآلي وبين الملخصات التي تم إعدادها مسبقا بواسطة مجموعة من الأشخاص، وبدون الأخذ في الاعتبار موقعها داخل النص. كلما زاد عدد التقاطعات المشتركة كلما اعتُبِرَ ذلك مؤشرًا لجودة التلخيص.
- الطريقة الثالثة: تعتمد على نظرية «كلود شانون» المعروفة باسم «نظرية المعلومات»، وتُستخدم في ضغط البيانات عند نقلها على شبكات الاتصال من مكان إلى آخر بغرض سرعة نقلها، ولكن مع القدرة على استرجاع البيانات الأصلية الكاملة من البيانات المضغوطة بعد استقبالها.
- الطريقة الرابعة: تعتمد على الاستعلام، وتعمل كالآتي:

يقوم مجموعة من الأشخاص بقراءة النص الأصلي للوثيقة المراد تلخيصها ثم يقومون بوضع مجموعة من الأسئلة تعكس أهم عناصر النص الأصلي.

تقوم مجموعة أخري من الأشخاص بالآتي:

- إجابة الأسئلة بدون الاطلاع على أي شيء (لا الوثيقة الأصلية ولا الملخص الآلي) ويطلق على هذه الإجابة مصطلح «خط الأساس».

- إجابة الأسئلة بعد الاطلاع على الملخص الآلي.
- إجابة الأسئلة بعد الاطلاع على الوثيقة الأصلية.

وتحتسب جودة التلخيص الآلي (نسبة الاحتفاظ بالمعلومة) بنسبة الإجابات الصحيحة التي أجابها المجموعة الثانية من الأشخاص بعد الاطلاع على الملخص الآلي مقارنة بالإجابات قبل وبعد الاطلاع على النص الأصلي للوثيقة.

• الطريقة الخامسة بالاعتماد على تصنيف نتائج الملخص الآلي، وتعمل كالتالي:

- يتم تجميع ٥٠٠٠ مقالة إخبارية من خمسة مجالات مختلفة (صحة، سياسة، ...) بواقع ١٠٠٠ مقالة لكل مجال.
- يقوم الملخص الآلي بتلخيص هذه المقالات الإخبارية.
- يقوم مجموعة من الأشخاص (بدون الاطلاع على المقالات الأصلية) بتصنيف الملخصات إلى المجالات المختلفة.
- يتم حساب نسبة أعداد الملخصات التي تم تصنيفها بصورة صحيحة متوافقة مع تصنيف أصل المقالة.
- يتم حساب نسبة أعداد الملخصات التي تم تصنيفها بصورة خاطئة مقارنة مع تصنيف أصل المقالة.
- يتم حساب جودة التلخيص (نسبة الاحتفاظ بالمعلومة) بدلالة النسب المحسوبة أعلاه.

في أغلب الطرق السابقة يستخدم مقياس (ROUGE) ومقياس (F–measure) للتعبير عن جودة التلخيص.

وبالطبع فإن التحدي الرئيسي لعمليات التلخيص هي الوصول إلى نسبة عالية من الاحتفاظ بالمعلومة، وفي نفس الوقت استخدام نسبة ضغط كبيرة.

ونأتي الآن إلي توضيح كيفيَّة إنجاز عمليات التلخيص الآلي.

٣- أساليب التلخيص الآلي

يمكن تقسيم أساليب التلخيص الآلي إلى المجموعات التالية:

١- أساليب إحصائية (أساليب تعلم الآلة).

٢- أساليب معالجة اللغة الطبيعية (على المستوى الصرفي والنحوي).

٣- أساليب المعالجة الدلالية وأساليب شبكات الكلمات.

٤- أساليب الحسابات المرنة، مثل: الشبكات العصبية، الخوارزمات الجينية، المنطق الفازي، وذكاء الأسراب.

٣,١- الأسلُوب الإحصائي لعملية التلخيص

يتصف الأسلُوب الإحصائي بأنه عند اختيار الجُمَل التي تُكَوِّنُ ملخص النص لا يتم النظر إلى أي تحليلات لغوية، مثل التحليل الصرفي أو النحوي أو الدلالي لمحتويات الوثيقة، ولكن يؤخذ في الاعتبار بعض أو كل العناصر التالية:

- الجُمَل التي تحتوي على كلمات ذات معدل تكراري عالٍ في النص.
- العبارات المميزة.
- الجُمَل التي تقع في عناوين الوثائق.
- الجمل التي تقع على رأس الفقرات والأجزاء داخل النص.
- موقع الجملة داخل النص.
- طول الجملة (عدد كلماتها).

وعادة تُعطى الأولوية للجُمَل التي تقع في عناوين الوثائق وللجُمل التي تحتوي على كلمات ذات معدل تكراري عالٍ في النص.

٣,٢- الأسلُوب اللغوي لعملية التلخيص

هنا يتم الأخذ في الاعتبار الخصائص اللغوية للنص المراد تلخيصه. ويتكون من ثلاث مراحل:

- التعرف على موضوع النص (Topic Identification).

- التفسير (Interpretation).
- توليد الملخص (Generation).

ونستعرض فيما يلي كل مرحلة من هذه المراحل:

مرحلة التعرف على موضوع النص

ويتم التعرف على موضوع النص بمجموعة من الطرق، منها: طريقة بنية الخطاب (Discourse Structure)، وطريقة التسلسل المعجميّ (Lexical Chains)، وهي الأكثر شيوعا الآن.

طريقة بنية الخطاب: ويقصد بها اكتشاف مجموعة الجمل التي تغطي سياق النص؛ ولتوضيح ذلك نفترض أنَّنا نُريدُ تلخيص النص التالي:

- توفير أجهزة رئيسية وأجهزة شخصية مع ملحقاتها (طابعات) وفقا للتقنيات الحديثة ومتطلبات العمل.

-تم استلام برامج النظم المالية بعد تطويرها ومراجعة تطبيقاتها للتأكد من مطابقتها لاحتياجات القطاع المالي من إجراءات وقواعد وتم التطبيق والاستخدام الفعلي لتلك النظم بجميع تطبيقاتها اعتبارا من 19/5/2004.

- توفير خدمة الانترنت للموظفين (جاري العمل على توفير خدمة E1 لتشغيل خدمة الانترنت من المنزل).

- قامت إدارة الشئون الإدارية بتزويد إدارة نظم المعلومات ببياناتها وذلك لوضعها على صفحة الانترنت. ميكنة أنظمة وبرامج الهيئة الإدارية والمالية.

- تم تركيب أجهزة اتصال وحماية (داخلية وخارجية) بالإضافة إلى الحماية من الفيروسات. ولم يتم استلام دليل المستخدم

- تم البدء في إدخال بيانات الموظفين من واقع ملفاتهم وبلغت نسبة النجاح (65%). خدمة الانترنت: كما تم عمل البنية التحتية لشبكة الحاسب الآلي وتوفير خدمة الاتصال بالانترنت.

ويتم استخلاص الجمل التي تغطي سياق النص، وهي الجمل التي تجيب عن الأسئلة من نوعية: ماذا حدث، لماذا حدث، كيف حدث، متى حدث، مَن فعل، وهكذا. فكلمات مثل: توفير، تزويد، إدخال، تشغيل، وفقا، اعتبارا من.. تساعد في تحديد الجمل التي تغطي سياق النص.

ويلعب التعرف على الكائِنات الاسميَّة، مثل أسماء الأشخاص وأسماء الجهات، دورًا كبير في إنجاح هذا الأسلُوب من التلخيص.

طريقة التسلسل المعجمي: يقصد بها استخلاص سلاسل الكلمات ذات الصلة في النص. وتهدف هذه الطريقة إلى التعرف على الموضوعات المحورية داخل النص. تم تطوير هذه الطريقة في أوائل التسعينيَّات بواسطة موريس وهاريس بناء على أبحاث سابقة في منتصف السبعينيَّات حولَ وظائف اللغة، للُّغَوِيّ المعروف مايكل هاليداي.

وتعتمد هذه الطريقة على مفهوم التماسك النصي الذي يربط الجمل بعضها ببعض من خلال أدوات لغوية مثل حروف الإشارة، الضمائر، الاستبدال، الحذف، الاقتران وغيرها.

بصورة أخرى، فإن السلاسل المعجمية تمثل التتبع لكائِن اسميٍّ محدد داخل النص، هذا الكائِن الاسمي يتم التعبير عنه بطرق مختلفة؛ فقد يأتي في صورة اسم علم ثم ضمير يعود عليه. ولا يشترط ذلك أن يكون في نفس الجملة ولا حتى نفس المقطع من النص.

وهنا تبدو مشكلة الالتباس؛ فتحديد «مَن يعود على مَن» ليس بالمسألة السهلة بالنسبة للحاسب حيث يتطلب فك الالتباس اللجوء إلى قواعد اللغة وإلى المعرفة العامة والتخصصية. وكذلك مشكلة الالتباس التي تنشأ نتيجة المعاني المتعددة للكلمة وكيفية اختيار المعنى الصحيح للكلمة وغير ذلك من التحديات اللغوية الكثيرة.

وبهذا فإن مرحلة التعرف على موضوعات النص تنتهي مع الوصول إلى مجموعة السلاسل اللغوية التي تم تحديدها في النص، ونأتي بعد ذلك إلى مرحلة التفسير.

مرحلة التفسير

يُقصَد بالتفسير في هذا السياق تخصيص سلسلة واحدة فقط وبالضبط لكل تواجد لكائِن اسمي في النص. هذه العملية تستغرق وقتًا طويلاً (يُعَبَّر عنه بالدَّالَّة الأسِّيَّة لعدد الأسماء الموجودة بالنص) إذا أخذنا في الاعتبار كَمَّ الاحتمالات الهائل للتفسير والوصول إلى التفسير الصحيح، أو على الأقل التفسير الأفضل. ويقصد بالأفضل هنا التفسير الذي يغطي أطول السلاسل المعجمية المستخرجة.

وللتغلب على مشكلة التعامل مع جميع الاحتمالات الممكنة فقد اقترحت رجينا بارزيلا، ومايكل الحداد وسيلبر ومككوي ربط السلسلة المعجمية بمفهوم معين وربط هذا المفهوم بمعنى مأخوذ من نظام ووردنت (WordNet)، وهي قاعدة بيانات معجمية

للغة الإنجليزية حيث يتم تجميع الكلمات الإنجليزية في مجموعات من المترادفات تدعى (synsets)، وتوفر تعريفات قصيرة عامة، وتسجل العلاقات الدلالية المختلفة بين مجموعات المترادفات المختلفة. كذلك تعتمد طرق «رجينا بارزيلا» و «مايكل الحداد» و«سيلبر» و «مككوي» على استخدام مُعرِبات سطحيّة بسيطة وأدوات لتعيين أقسام الكلام (مُعَنوِنات) للكلمات للتعرف على الأسماء. وتأخذ طريقة «سيلبر» و «مككوي» زمنًا خطيًّا (بدل الزمن الأسِّيّ الطويل) بدلالة عدد الأسماء في النص للوصول إلى أحسن التفسيرات للسلاسل المعجمية.

تنتهي مرحلة التفسير بتحديد أهم (أفضل، أقوى) السلاسل المعجمية في النص بناء على أسلُوب تقييم يأخذ في الاعتبار عدد المرَّات الَّتي يتكرَّرُ فيها الكائِن الاسميّ وعلاقاته السابقة مع باقي كلمات السلسلة. ومن الجدير بالذكر أنَّ هناك أساليب كثيرة لتقييم السلاسل المعجمية المستخرجة من النص. فعلى سبيل المثال، يمكن الأخذ في الاعتبار عناصر غير لغوية، مثل حجم ولون وموقع كتابة الجمل داخل النص كدلالة لأهمية السلسلة. بعد ذلك تأتي مرحلة توليد الملخص.

مرحلة توليد الملخص (Generation)

بعد تحديد أهم (أفضل، أقوى) السلاسل المعجمية في النص يتم اختيار جملة واحدة من كل منها؛ ولكن أي جملة يتم اختيارها؟

إحدى البدائل لكل سلسلة قوية أن يتمّ اختيار أول جملة تشير إليها وتضمينها (بالترتيب) في الملخص. بديل آخر لكل سلسلة قوية بأن يتم اختيار أول جملة تشمل الممثل الاسمي الذي يعبر وترتبط به السلسلة وتضمينها (بالترتيب) في الملخص، مع ملاحظة أن الممثل الاسمي هو المعنى المناظر المأخوذ من شبكة الكلمات (WordNet)، والذي يعبر وترتبط به السلسلة.

للشرح التفصيلي لاستخدام السلاسل المعجمية في التلخيص يُفَضَّل الرجوع إلى المقالات المرجعية التَّالية:

1. "Lexical cohesion computed by thesaural relations as an indicator of the structure of text" by Morris, J. & G. Hirst, 1991.
2. "Using Lexical Chains for Text Summarization", by Regina Barzilay & Michael ElHadad, 1997.

3. "Efficiently Computed Lexical Chains as an Intermediate Representation for Automatic Text Summarization", by Silber, G. & K. McCoy, 2002.

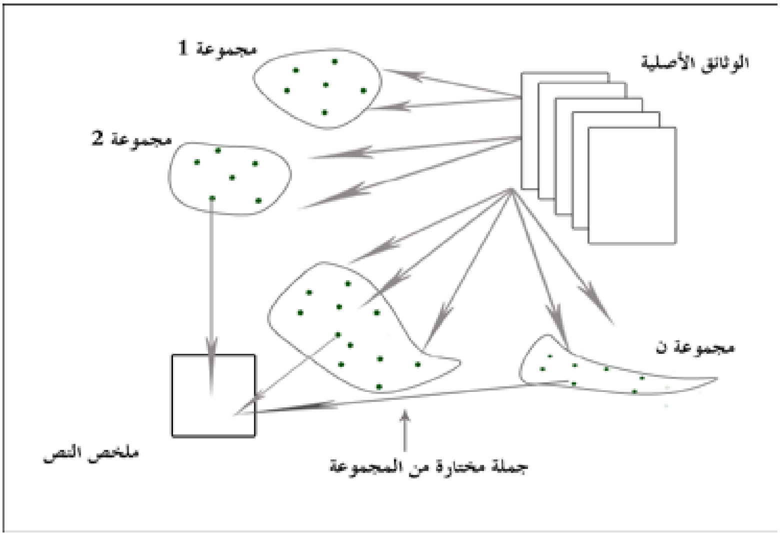
٣,٣- أسلُوب الجمع بين تعلم الآلة والمعالجة اللغوية

في هذا الأسلُوب تتم عملية التلخيص في ثلاث مراحل:

- قراءة الوثائق.
- فهم الوثائق من خلال بناء التمثيل الدلالي لمحتويات الوثيقة.
- توليد الملخص من هذا التمثيل.

ونظرًا لأن عملية فهم وتوليد التمثيل الدلالي الغني للنص معقدة للغاية وليست ممكنة حتى الآن، فإن معظم نظم التلخيص تكون من نوع استخلاص عدد محدود من الجمل من النص الأصلي مع إعادة صياغة هذه الجمل لحذف الحواشي منها. ولكن يعيب هذه الطريقة أن الملخصات المستخرجة من النص تكون غير متماسكة في العادة، ولكن يميزها أنها غير مكلفة الحل ولا تتطلّب أنطولوجيات معاونة.

يعمل أسلُوب الاستخلاص كما هو موضح بالشكل التالي:



الشَّكل ٤-٦: كيفية عمل الملخص الآلى للنصوص باسلوب الاستخلاص.

- بعد قراءة الوثائق الأصلية، يتم تقسيم الجمل الموجودة إلى مجموعات من الجمل المتشابهة، ويطلق على هذه الخطوة «عملية التجميع» (Clustering).
- يتم ترتيب المجموعات الناتجة وهي في العادة كبيرة العدد، حيث يعكس الترتيب أهمية المجموعة.
- يتم اختيار جملة واحدة من المجموعات الأولى حسب ترتيب المجموعات.
- في العادة تحتوي الجمل المختارة على حواشٍ وتكرارات يمكن الاستغناء عنها، لذا يتم إعادة صياغة الجمل من خلال حذف هذه الحواشي، وهنا يأتي دور اللغة في إعادة صياغة الجملة.

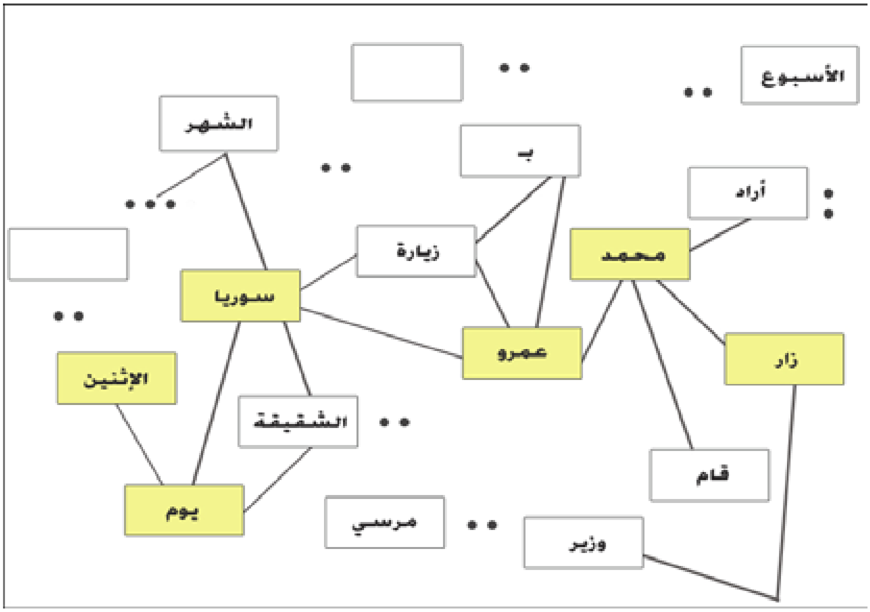
٣,٤- أسلُوب شبكات الكلمات

يُستخدم هذا الأسلُوب، المقترح بواسطة «كاتجا فيلبوفا» - من جوجل - في دمج مجموعة من الجمل في نفس الموضوع إلى جملة واحدة. ويعتمد على بناء شبكة بين كلمات مجموعة من الجمل المراد اختصارها حيث تُمَثَّل كل كلمة بعقدة داخل الشبكة، والخطوط التي تصل الكلمات تمثل تتابع الكلمات داخل كل جملة، مع ملاحظة أن الكلمات المكررة من الممكن دمجها في عقدة واحد في ظل شروط معينة.

يبدأ بناء الشبكة للجملة الأولى في صورة سلسلة من العقد بواقع عقدة لكل كلمة في الجملة. عند التعامل مع جملة جديدة يتم بناء عُقَد جديدة لكلماتها أو تدمج كلماتها مع كلمات الجمل السابقة إذا لم يكُن هناك التباسٌ في المعنى. ويتم بناء عقد لحروف الجر وأسماء الإشارة أو إهمالها تحت شروط معينة.

مثال لشبكة كلمات تناظر أربع جمل مختلفة عن نفس الموضوع:

- أراد محمد عمرو زيارة سوريا الشهر الماضي لكنه أجل خططه حتى يوم الاثنين الماضي.
- قام محمد عمرو بزيارة دولة سوريا الشقيقة يوم الاثنين.
- زار محمد عمرو نائبا عن الرئيس محمد مرسي سوريا يوم الاثنين الماضي.
- الأسبوع الماضي زار وزير الخارجية السيد عمرو المسؤولين السوريين.



الشَّكل ٤-٧: يوضح أسلوبًا لشبكات الكلمات

وهذا الأسلُوب في الدمج يعتمد بصورة كبيرة على المعالجة اللغوية والتعرف على مواطن الالتباس في الجملة والتعرف على حروف الربط والإشارة وربطها بمدلولها.

٣,٥- أساليب الحسابات المرنة لتلخيص النصوص

يندرج تحت هذه الفئة عدد من التقنيات التي تحاول أن تحاكي الكائنات الحية في التفكير أو في التطور أو تتسم بالغُمُوض في التعبير.

فمن التقنيات التي تحاكي الكائنات الحية في التفكير نجد شجرة القرارات والشبكات العصبية وذكاء الأسراب، ومن التقنيات التي تحاكي الكائنات الحية في التطور نجد الخوارزمات الجينية، ومن التقنيات التي تتسم بالغُمُوض نجد المنطق الفازي.

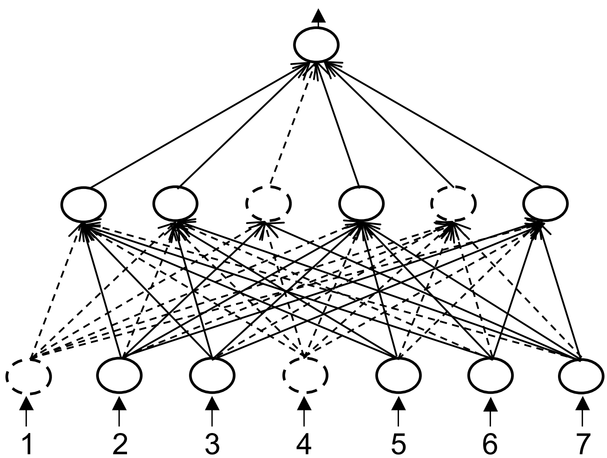
وتعتمد معظم هذه الأساليب على إيجاد مجموعة من الخصائص لكل جملة ثم استخدامها كبارومترات يتم تحديد مدى أهمية الجملة للاحتفاظ بها في ناتج التلخيص.

والخصائص التالية تعد الأكثر انتشارا من حيث الاستخدام في هذه الأساليب:

- موقع الجملة في النص.

- مدى تشابه الجملة مع عنوان النص.
- مدى محورية الجملة (ويمكن قياسها بمدى احتوائها للكلمات الأكثر تكرارا في النص أو بطرق أخرى).
- مدى احتواء الجملة على كلمات إيجابية مثل كلمات سعيد، حيوي، نشط.
- مدى احتواء الجملة على كلمات سلبية مثل فقير، مرهق.
- مدى احتواء الجملة على كيانات اسمية، مثل رئيس الدولة.
- مدى احتواء الجملة على بيانات عددية.
- طول الجملة مقارنة بباقي الجمل في النص.
- ...

ففي حالة الشبكات العصبية، على سبيل المثال، يتم بناء الشبكة من ثلاث طبقات. الطبقة الأولى تتكون من مجموعة عقد تحمل قيم الخصائص المختارة للجملة (خ١، خ٢، ...، خ٧، ...). الطبقة الثانية تتكون من مجموعة من العقد ويطلق عليها الطبقة المخفية، وعدد العقد بها أقل من عدد عقد الطبقة الأولى ويتم التوصل إليها بالتجربة والخطأ. الطبقة الثالثة والأخيرة مكونة من عقدة واحدة وهي التي تحدد مدى اختيار الجملة ضمن النص المراد تلخيصه.



الشَّكل ٤-٨: شبكة عصبية لها ٧ مدخلات تحدد خصائص الجملة ولها مخرج واحد يحدد مدى أهمية الجملة

والعلاقات والدالات الحسابية التي تربط قيم العقد في الطبقات المختلفة يتم تعلمها من خلال المئات بل الآلاف من أمثلة التلخيص اليدوي.

وكما ذُكِرَ آنفًا.. فإن هذه التقنيات والأساليب المستخدمة في تلخيص النصوص تستخدم أيضا في باقي مجالات التَّنقيب في النُّصُوص، ومجالات أخرى، مثل التَّنقيب في البيانات ومجالات التعرف على الصور والتعرف على الكلام والتعرف على الكتابة.

٤- نماذج من أنظمة التلخيص الآلي

يوجد الآن العديد من أنظمة التلخيص على المستوى التجاري، ولكن معظمها يخدم اللغة الإنجليزية وقليل منها يخدم اللغة العربية. وجدير بالذكر أنه لا يزال هناك الكثير من البحث والجهد المطلوبَين لرفع جودة هذه الأنظمة، وبالأخص بالنسبة لتلخيص النصوص العربية.

٤,١- نماذج من أنظمة التلخيص للنصوص الأجنبية

- **(SweSum)**

هو أول نظام لتلخيص النَّصّ للغة السويدية. وهو يلخص نصوص الأنباء السويدية المكتوبة بتنسيق (HTML) على شبكات الإنترنت. ناتج التلخيص عبارة عن عدد من ١٠- ٥٠ من الكلمات الحاكمة. وتتراوح دقة التلخيص للنصوص الصحفية من ٤٠٪ إلى ٨٤٪ وذلك للنص الأصلي الذي يصل طوله في المتوسط إلى ١٨١ كلمة. ونظام (SweSum) متاح أيضا للغات الدنمركية والفارسية والنرويجية والإنجليزية والإسبانية والفرنسية والإيطالية واليونانية والألمانية.

ويستند (SweSum) على الأساليب الإحصائية واللغوية وأساليب الذكاء الاصطناعي. وتتمّ عملية التلخيص واختيار الكلمات الرئيسية من خلال قيام النظام بحساب تكرار الكلمات الرئيسية في النص الصحفي وموقع هذه الجمل في النص. ويأخذ في الاعتبار حجم حروف الكتابة لهذه الكلمات، وهل هي موجودة بالفقرات الأولى في النص أم لا، وما إذا كانت القِيم الموسومة قيمًا عددية أم لا.

- **(SUMMARIST)**

هو محاولة لتطوير تكنولوجيا قوية للتلخيص بأسلُوب الاستخراج للإفادة في مجالات البحث العلمي، ومن ثم مواصلة البحث وتطوير تقنيات الوصول للنظرة التجريدية للنص. هذا العمل يوائم بين عمق التلخيص مقابل متانة التلخيص (ويقصد بالمتانة القدرة على التعامل مع النصوص التي تشتمل على أخطاء لغوية). فيكون التركيز على استخدام أنظمة تحليل وتفسير المدخلات بصورة عميقة بما يكفي لإنتاج ملخصات جيدة، أو على النصوص المقيدة بصورة أو بأخرى ولكن لا يمكن تحليلها بطريقة عميقة بما يكفي لصهر المدخلات بصورة صحيحة، وبالتالي تؤدي فقط إلى استخراج موضوع النص.

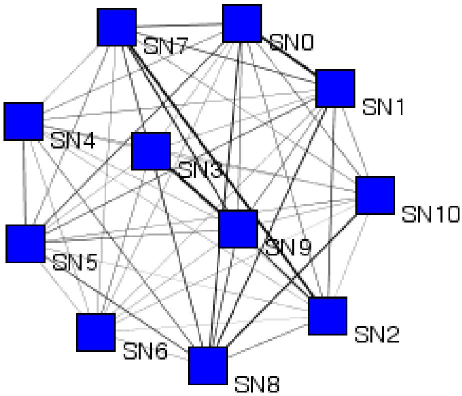
حتى الآن، ينتج نظام (SUMMARIST) ملخصات الاستخراج في خمس لغات (حيث تم ربطه بمحركات الترجمة لهذه اللغات في نظام MuST للترجمة الآلية). العمل جار على حد سواء لتوسيع قدرات النظام في التلخيص القائم على استخراج الجمل المهِمَّة في النص أو لبناء مجموعة كبيرة من قواعد المعرفة المطلوبة لعمليات التلخيص القائمة على الاستدلال والتجريد للنص الأصلي.

- **(LexRank Summarizer)**

LexRank هو نظام لتلخيص النصوص الإنجليزية تم تطويره في جامعة ميتشجان الأمريكية، وهو متاح للعمل على شبكة الإنترنت عبرَ الموقع:

http: / /clair.si.umich.edu /clair /lexrank

يمكن للمستخدم كتابة النص المراد تلخيصه أو تحميل النص من ملف سبق إعداده. ويعتمد النظام على الأساليب الإحصائية والرسوم الشبكية. يقوم النظام بتحويل الجمل النصـية إلى متجهات عددية ثم يحسب قيمة الزاوية بينها (Similarity Cosine) (وهي إحدى طرق حساب التشابهات بين جمل النص) ثم يقوم بعد ذلك بحساب مصفوفة الجوار لجميع الجمل الموجودة بالنص. ويأتي ناتج التلخيص من خلال الاحتفاظ بالجمل ذات أعلى قيم بداخل هذه المصفوفة. الشكل التالي يعبر عن علاقة الجوار والتشابه في نص مكون من إحدى عشرة جملة يرمز لها داخل الرسم بالرموز SN0, SN1, SN3, … , SN10.



الشَّكل ٤-٩: مصفوفة جوار مُعَبَّر عنها برسم شبكي

- **(Intellexer Summarizer)**

وهو نظام للتلخيص الآلي، يستخدم أساليب معالجة اللغات الطبيعية بكثافة، فيقوم بالتحليل الصرفي والنحوي والدلالي للنص كجزء متكامل. وهو متاح للاستخدام من خلال الموقع التالي على شبكة الإنترنت:

http: / /www.fileguru.com /Intellexer-Summarizer-SDK /info

- **بعض الأنظمة التي تعمل من خلال شبكة الإنترنت:**

- **Automatic Text Summarizer**
  http / /:www.makeuseof.com /dir /automatic-text-summarizer-text-summarization-tool /
- **The Open Text Summarizer**
  http: / /libots.sourceforge.net /
- **Kify Online Text Summarizer**
  http: / /text.kify.com /
- **Intellexer Summarizer 3.1**
  http: / /summarizer.intellexer.com /
- **PERTINENCE SUMMARIZER**
  http://www.pertinence.net/ps/summarizer_url.jsp?ui.lang=en

- **QuickJist summarizer 1.2**
  http://www.filecluster.com/Internet/Browser-Tools/Download-QuickJist-summarizer.html
- **Sinope Summarizer**
  http://www.sinope.info/en/Download
- **Copernic Summarizer**
  http://www.copernic.com/en/products/summarizer/

٤,٢- نماذج من أنظمة التلخيص للنصوص العربية

- **نظام لخّص (Lakhas)**

(Lakhas) هو نظام للتلخيص الآلي تم تطويره بجامعة مونتريال الكندية. يستخدم النظام الأسلُوب الإحصائي في عمليات التلخيص التي تتم على المراحل التالية: تجزئة النص إلى مجموعة من الجمل، تجزئة الجمل إلى كلمات، وضع حروف الكلمات في صور موحدة (مثل هـ، ة ومثل أ، ا)، إزالة كلمات الوقف (stop words) (مثل: الذي، التي) ثم حذف السوابق واللواحق من الكلمة (Affixes) (وعدم اللجوء إلى استخلاص جذور الكلمات) ثم حساب المعدل التكراري للكلمات المستخلصة، ثم حساب الوزن النسبي لكل جملة معتمدا على معدل تكرار كلماتها وعلى موقع الجملة في النص، وأخيرا يتم استخلاص الجمل ذات الوزن النسبي العالي لتكوين الملخص المطلوب.

- **نظام (ACBTSS)**

(Arabic Concept-Based Text Summarization System)

يعتمد هذا النظام (من جامعة إسكس Essex البريطانية) على تقنية بايز الإحصائية وتقنية البرمجة الجينية حيث تُستَخدمان في أنظمة تصنيف النصوص. يحتاج هذا النظام إلى مُدَوَّنة لغويَّة مُرَمَّزة ومُزَوَّدة بالحواشي، تُستَخدم في تدريب النظام على استخراج خصائص الجمل التي يتم الاحتفاظ بها في ناتج التلخيص. وهذه الخصائص يتم تحديدها من خلال المعالجة اللغوية للنص (تحليل صرفي، ترميز أجزاء الكلام) ومن خلال موقع الجمل داخل النص بالإضافة إلى المعدل التكراري لكلمات كل جملة داخل النص.

- **نظام (The Summarizer of AramediA)**

يتميز هذا النظام (من شركة صخر) بوجود وظيفة تصحيح الأخطاء اللغوية الشائعة ثم يتم التلخيص من خلال استخدام أساليب إحصائية وتحريرية ولغوية للتعرف على أشباه الجمل الحاكمة في النص (الكلمات المفتاحية). ويستخدم النظام في تلخيص النصوص الإنجليزية والعربية.

http: / /aramedia.com /summarization.htm

٥- الخلاصة

يعتمد بناء أنظمة تلخيص عالي الجودة على أساليب معالجة اللغات الطبيعية مع التقنيات الحديثة في مجال تعلم الآلة والذكاء الاصطناعي، ولازال هناك تحديات بحثية وتطبيقية كثيرة في هذا المجال.

أهم هذه التحديات الآتي:

١- ما هي الميزات المهمَّة لنظم تلخيص النص والتي تعتمد على استخراج الأفكار الرئيسية من النص الأصلي للوثائق؟

٢- كيف يمكن التعامل مع الجمل الغامضة في النص الأصلي للوثائق، إن وجدت؟

٣- كيف نستطيع أن نقيِّم نظم تلخيص النص؟

ومن سمات الاتجاهات الحديثة في هذا المجال تحول الاهتمام من تلخيص النصوص العلمية والإخبارية إلى مراجعة واستعراض المنتجات المتاحة عبر الإنترنت، مثل المقالات الطبية الحيوية، وتتبع موضوعات التعليم، وتتبع رسائل البريد الإلكتروني، وتتبع المدونات على الإنترنت.

ومن أهم الاتجاهات البحثية الدمج بين أكثر من أسلُوب تقني مع الاهتمام بالخصائص الدلالية لكلمات الجمل ومكوناتها.

المبحث الثَّالث

استنباط اتِّجاهات الرَّأي العامّ

١ – أهمية تنقيب الآراء.

٢ – مهام وأساليب التنقيب عن الآراء.

٣– التنقيب في الآراء واللغة العربية.

٤ – الموارِد اللغوية اللازمة المتاحة والمطلوبة.

٥ – التوجهات المستقبلية والتحديات الَّتي تواجه تنقيب الآراء.

استنباط اتجاهات الرأي العام
(متابعة تطور الآراء على الشبكات الاجتماعية)

يُعتبرُ فهم اللغات الطبيعية واحدًا من أكبر تحديات الذكاء الاصطناعيّ أو هو المشكلة الكاملة في الذكاء الاصطناعيّ؛ ويُمَثِّل استخلاص الرأي من بين النصوص والتعرف على أجزاء النص الَّتي تحتوى على آراء مشكلةً تتصل بمعالجة اللغات الطبيعية.

التَّنقيب في الآراء (Opinion Mining) أو استخلاص الآراء أو وجهات النظر (sentiment extraction /Opinion) أو تحليل وجهات النظر (Sentiment Analysis) هي مرادفات تتصل بنفس المعنى.

تنقيب الآراء هو مجال البحوث الَّتي تسعى إلى تمكين النظم الآلية من تحديد الآراء البشرية من النصوص المكتوبة (أو المنطوقة مع التطور) بلغة بشرية طبيعية، وهو يتعقب ويبحث في تحديد وجهات النظر الَّتي تقع ضمن النص.

تنقيب الآراء هو: استخراج الآراء الواردة في النصوص، أو هو علم يقوم بدراسة استخراج الآراء باستخدام تقنيات استرجاع المعلومات IR، والذكاء الاصطناعيّ AI، ومعالجة اللغة الطبيعية NLP.

يتعلق المجال أيضا ويرتبط ارتباطا وثيقا بتلخيص الآراء من المحتوى المقدم من المستخدمين أو إعلام ما ينتجه المستخدمون على الإنترنت، أو ما يُعرض في المنتديات ومجموعات النقاش والمدونات والشبكات الاجتماعية، وتصنيف تلك الآراء (Sentiment classification) واستعراضها وتحليلها وكشفها.

ينسحب تنقيب الآراء على حوسبة اللُّغة، واسترجاع المعلومات IR، وتنقيب النصوص، ومعالجة اللغات الطبيعية، وتعلم الآلة، والإحصاء، والتحليل التنبؤي؛ وهناك العديد من التقنيات التي يمكنها إنجاز هذه المهام.

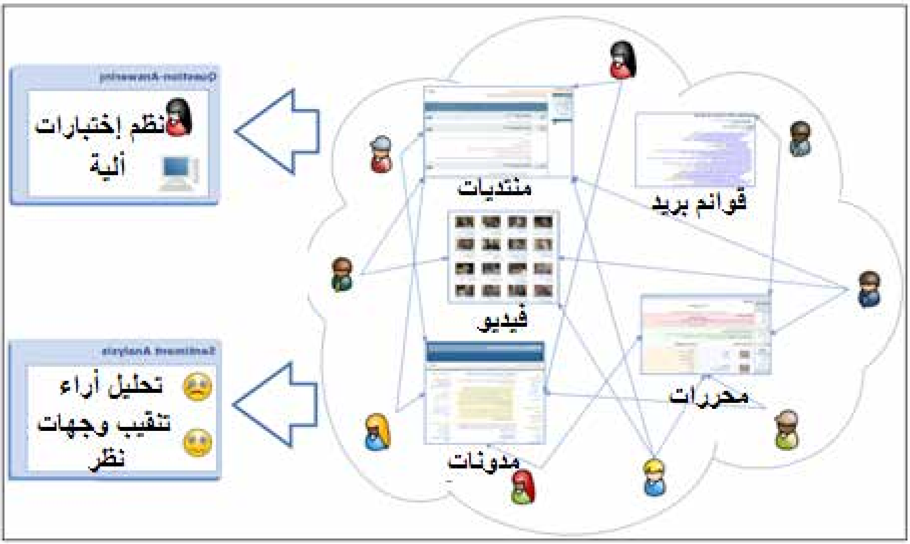
١ – أهمية تنقيب الآراء

في دراسة مسحية حديثة عن تطبيقات تقنيات التنقيب في النصوص تم الإشارة إلى أن ما يقرب من ٨٠ بالمئة من المعلومات المتاحة على الإنترنت مخزنة في شكل نصوص (بالمقارنة بالبيانات الرقمية) وإنه ومن الصعب استخلاص معلومات معينة منها.

ولما كانت الغالبية العظمى من المخزون الطبيعيّ مسجلا في كلمات ونصوص، فإن هناك نوعَين أساسيَّين من المعلومات النَّصِّيَّة، هما: الحقائق، والآراء. وتعمل معظم تقنيات معالجة المعلومات الحالية (مثل آلات البحث) على الحقائق (بافتراض صحتها)، ويمكن التعبير عن الحقائق بكلمات أساسية (كلمات حاكمة) تعبر عن الموضوع. من أمثلة الحقائق: بيانات تأسيس حزب، بيانات عن التوزيع السكاني لدولة ما، ما آخر مقالة للكاتب نجيب محفوظ؟، ما أعلى قمة جبل في العالم؟ وأين تقع؟، ما هي الدول الأعضاء في منظمة «يونسكو»؟ ومن هو مديرها الحالي؟، وهكذا.

الحقائق ذات أهمية كبيرة في الحياة الواقعية الحقيقية إلا أن الآراء أيضا تلعب دورا حيويا وأساسيا لمعرفة ما يجرى وردود الأفعال.

تم إنتاج طائفة واسعة من التطبيقات الَّتي تتيح تنقيب وجهات النظر (شكل ٤–١٠)، وقامت بالتركيز على قدر كبير من البحوث في السنوات الأخيرة، وقد تم التوصل إلى دقة عالية في التصنيف باستخدام مجموعة متنوعة من التقنيات، يعتمد معظمها اعتمادا كبيرا على علوم الإحصاء والذكاء الاصطناعي وتعلم الآلة وعلوم معالجة اللغات الطبيعية.



الشَّكل ٤-١٠: طائفة واسعة من التطبيقات الَّتي تتيح تنقيب وجهات النظر.

أصبحت أتمتة استخراج الآراء من النص مجالا يحظى باهتمام متزايد، ونظرا للكمية المتزايدة من المحتوى المقدم من المستخدمين والمتاحة على الشبكة فقد ازدادت أهمية قدرة دقة قياس الآراء بتطبيقات عملية أكثر من أي وقت مضى.

والإحصائية التالية لها دلالة عن أهمية هذا العلم الحديث، فوفقا لاثنتَين من الدراسات الاستبيانية لأكثر من ٢٠٠٠ من الأمريكيين البالغين لكل منهما:

١- تبين قيام ٨١٪ من مستخدمي الإنترنت (أو ٦٠٪ من الأمريكيين) بعمليات البحث على الإنترنت بما لا يقل عن مرة واحدة على منتج ما (منتجات مثل: أدوات التجميل، أجهزة المحمول، ...).

٢- تبين أنه من ٧٣٪ إلى ٨٧٪ من بين الذين يقرؤون التقارير والمقالات ووجهات النظر المنشورة على الإنترنت عن المنتجات والمطاعم والفنادق والخدمات المختلفة (مثل وكالات السفر أو الأطباء) قد أفادوا أن هذه التقارير لعبت دورا كبيرا على قرارهم في شرائهم هذه المنتجات أو ذهابهم إلى هذه الأماكن.

٣- تبين أن ٣٢٪ من الذين شاركوا في الاستبيان قد قاموا بالتعبير عن رأيهم في تصنيف جودة منتج أو خدمة أو شخص، وذلك باستخدام نظم التقييم الآنية

المتاحة على الإنترنت، وأن ٣٠٪ (من بينهم ١٨٪ من كبار السن) قد نشروا تعليقا على الإنترنت أو قاموا بمراجعة بشأن منتج أو خدمة من الخدمات التي يُرَوَّج لها على الإنترنت.

يفيد تنقيب الآراء في العديد من التطبيقات، مثل:

١- المؤسسات والمنظمات من أجل تقييم المنتجات والخدمات.

٢- استخبارات السوق (Market intelligence).

٣- توفير المال والجهد، ومعرفة آراء ووجهات نظر المستهلكين.

٤- يساعد في معرفة الأفراد للمنتجات الَّتي تلقى اهتماما من الآخرين ووجهات نظرهم بشأنها.

٥- توفر نظم تنقيب النصوص للمؤسسات والشركات معلومات تنافسية من خلال معالجة كمية كبيرة من النصوص والحصول على الفوائد منها.

٦- تحليل ملفات العملاء، تحليل اتجاهات، ترشيح وتوجيه المعلومات، تتبع الأحداث، تصنيف الموضوعات الإخبارية، بحث ويب،... إلخ.

٧- يقوم تحليل ملفات العملاء بالتنقيب في البريد وشكاوى العملاء والتغذية المرتدة منهم، كما يمكن تحليل ملفات المرضى للحصول على اتجاهات مرضية وشكاوى وجودة الخدمات، وتحليل بث المعلومات وتنظيم وتلخيص اتجاهات الأخبار والتقارير، وتنقيب مستندات تخطيط موارد المؤسسة.

٨- التطبيقات كمكونات تكنولوجية فرعية: نظم التوصيات، التلخيص، إجابة التساؤلات.

٩- تطبيقات في الأعمال: استخبارات السوق، تحسين المنتج والخدمات.

١٠- فهم رأي المستهلك الذي يعبر عن صوته في الاتصالات اليومية.

١١- تطبيقات سياسية: كما هو معروف، يلعب الرأي دورا كبيرا في السياسة، وتركز بعض التطبيقات على فهم ما يفكر فيه المواطنون عند التصويت في الانتخابات أو تشريع القوانين.

١٢- تحليل المدونات: إنجاز تصنيف الاستقطاب والتصنيف الموضوعيّ على محتويات المدونات والمراسلات الَّتي تتم فيها.

١٣- اكتشاف الاختلافات في أنماط الحالة المزاجية على مدار الوقت (الخوف، الإثارة، الحزن، التعاطف، القلق،.. إلخ) الذي يظهر على نطاق واسع.

١٤- استخدام ربط المعلومات الزمنية لنمذجة الثقة والتأثير في نطـاق المدونات.

١٥- تحليل وجهات النظر في المدونات عن أعمال فنية وإبداعية وأفلام ومبيعات.

١٦- تفاعل الحاسب والإنسان، وتفاعل الإنسان مع الروبوت.

١٧- التعليم والامتحانات.

٢- مهام وأساليب التنقيب عن الآراء

يُعَبَّر عن الآراء داخل النصوص بإحدى وسيلتين. الوسيلة الأولى: وهي التعبير عن الرأي المباشر، مثل « تصميم هذه السيارة رائع»، والوسيلة الأخرى: من خلال التعبير المقارن، مثل «تصميم هذه السيارة أفضل من تصميم السيارة الأخرى». وواجب تقنيات التنقيب التعامل مع الآراء بأنواعها المختلفة.

وإذا كانت محركات البحث في صفحات الإنترنت تلبي بصورة أو أخرى حاجات المستخدم في البحث عن الحقائق من خلال استخدام الكلمات الحاكمة للتعبير عن متطلباته، فإن على محركات التنقيب في الرأي أن تلبي حاجات المستخدم في معالجة وإجابة أنواع الأسئلة التالية الخاصة بالرأي:

- ما هو رأي شخص أو جهة أو كائِن اسميّ معين في منتج أو كائن معين أو في إحدى خصائصه، مثل:

  ما رأي عباس العقاد في الكاتبة ميّ زيادة؟

- من هم الأشخاص أو الأعضاء ذوو الرأي المعين في موضوع أو منتج أو كائن معين أو في إحدى خصائصه. مثل:

  ما هي الدول الأعضاء في مجلس الأمن التي تتعاطف مع القضية الفلسطينية؟

- ما هو الرأي الإيجابيّ (أو الرأي السلبي) في منتج أو كائن معين. مثل:

ما هي مميزات الحاسب المحمول؟ ما هي عيوب بطاريات الحاسبات المحمولة؟

- مقارنة بين الرأي في أشخاص أو جهات أو كائِنات اسميَّة معينة أو في منتجات أو كائنات معينة. مثل:

ماذا يميز التِّلفاز من نَوع «إل سي دي» عن التِّلفاز العاديّ؟

- ما هو الرأي في منتج أو كائن معين؟ هنا لا يكتفي محرك التنقيب برأي شخص واحد أو جهة واحدة، إنما يجب أن يأخذ في الاعتبار الآراء المختلفة ومن المفيد أن تكون الإجابة طبقا لآراء الأغلبية مع التنويه عن النسبة. مثل:

ما رأي الجمهور في أداء الفريق القومي أمس؟

لكي تكون محركات التنقيب قادرة علــى التعامل مع الأنواع المختلفة من الأسئلة السابقة، ينبغي أن تتعاملَ مع المحتويات المختلفة لمكونات النص على النحو التالي:

- التعامل على مستوى عبارة داخل الجملة للتعرف على الكائِن (شخص، جهة، منتج، إلخ).
- التعامل على مستوى عبارة داخل الجملة للتعرف على خاصية من خصائص الكائِن واستخلاصها (درجة حرارة الغرفة، سعة ذاكرة الحاسب، تصميم السيارة، إلخ).
- التعامل على مستوى الجملة للوصول للرأي.
- التعامل على مستوى الوثيقة للوصول إلى تصنيفات الرأي المستخلصة من الجمل.
- أحيانا تحتوي الجملة الواحدة على أكثر من رأي أو مقارنة بين رأي وأخر مثل: «محمد يحب كرة القدم، ولكن عادل لا يكترث».

بالنظر إلى ما سبق يمكن أن نخلص إلى أن مفهوم الرأي يحتوى على ثلاث مكونات رئيسة، هي:

- صاحب الرأي أو حائز الرأي.

- الكيان أو الشيء موضوع الرأي (أي الذي نبحث عن الرأي عنه).
- الرأي ذاته.

لكي تستطيع محركات التنقيب عن الآراء الوصول إلى هذه المكونات الثلاثة من الجملة أو الوثيقة فإنها تقوم بمجموعة من المهام والوظائف المتعددة، مثل التعرف علي الكلمات والجمل اللغوية داخل المقال، التعامل مع المترادفات والمتَضادَّات، التعامل مع التطابق والجناس والتعامل مع الكلمات التي تحمل معنى الرأي والتعرف على الدلالات التعبيرية لهذه الكلمات والجمل، ثم تحديد وتصنيف رأي المقال. وعموما فإن قائمة المهام التالية تمثل حجر الزاوية في نظم التنقيب عن الآراء، والتي تأخذ معالجة وخصائص اللغات الطبيعية في الاعتبار:

- التعرف علي الكائِنات الاسميَّة.
- التحليل الصرفي والإعرابي للنص.
- بناء واستخدام قاعدة بيانات الدلالة المعجمية المعنية بالمشاعر (Sentiment Lexical Semantics Database).
- بناء واستخدام مُدَوَّنة نَصِّيَّة مُعَنوَنة بالتَّوصِيفات الدَّالَّة على الرأي (Opinion Annotated Corpora).
- التعرف على القائم بإبداء الرأي (ويطلق عليه اسم حائز الرأي) والتعرف على موضوع الرأي.
- التعرف على طبيعة الكلمات (كلمات موضوعية مُقارَنة بالكلمات التقديرية).
- تحليل المعنى التقديري للكلمة (Subjectivity Analysis).
- استخراج الرأي وتصنيف النص طبقا لذلك.
- تلخيص وجهات النظر المختلفة (Views summarization) (وتلعب دورًا كبيرًا في حالة تعدد الوثائق عن نفس الموضوع، سواء أكانت مكتوبة بلغة واحدة أم بلغات متعددة).

وفيما يلي سنلقي الضوء على أساليب تنفيذ بعض هذه المهام.

٢ , ١ - التعرف على أسماء الكائِنات (Recognition Entities Named)

توجد تقنيات متعددة للتعرف على الكائِنات الاسمية داخل النص. وتنقسم هذه التقنيات إلى الأنواع التالية:

- تقنيات مبنية على القواعد النَّحويَّة. وهي عادة عبارة عن قواعد مصاغة يدويًّا، تأخذ في الاعتبار الخبرة اللغوية العالية. وهذه التقنيات تعطي جودة عالية في التعرف على الكائِنات الاسميَّة، ولكنها في الغالب لا تغطي معظم الحالات، بالإضافة إلى إنها عالية التكلفة في الإعداد وتحتاج إلى أشهر من العمل من قبل اللغويين ذوي الخبرة الحسابية.
- تقنيات مبنية على النماذج الإحصائية للغة. وتتطلب عادة إعداد كمية كبيرة من النصوص التي يتم إضافة الحواشي إليها وتمييز الكائِنات الاسميَّة بينها يدويًّا. ويبقى دور برمجيات تعلم الآلة لاستخلاص وصياغة نماذج التعرف على الكائِنات الاسميَّة. وهي أيضا مكلفة الإعداد ولكن لا تحتاج خبرة اللغويين بمثل احتياج التقنيات السابقة.
- تقنيات مبنية على قوائم بالأسماء السابق إعدادها يدويا (أو قواعد بيانات متخصصة للكائِنات الاسمية Gazetteers)، وتعمل بنجاح في المجالات ذات الطبيعة التخصصية.

٢ , ٢- التعرف على القائم بإبداء الرأي (ويطلق عليه اسم «حائز الرأي»)

جذبت مهمة التعرف على «حائز الرأي» عددًا كبيرًا من الباحثين. وقد استعيرت تقنيات كثيرة من مجالات متعددة، مثل التعرف على الأصوات والتنقيب في البيانات العددية لتخدم هذه المهمة. وبدون الدخول في التفاصيل الفنية، نميز من بين التقنيات الكثيرة المستخدمة في التعرف على حائز الرأي التقنيات التالية:

- تقنيات نماذج ماركوف المخفيَّة (HMM– Models Markov Hidden).
- تقنيات الحقول الشرطية العشوائية (Fields Random Conditional).

- المنهج القائم على قواعد المعرفة.
- تقنيات التجميع والتصنيف (مثلا باستخدام حالة الفوضى القُصوى).
- منهج يعتمد على وجود معجم.
- تقنيات تعتمد على وجود معلم، مثل:
  - تقنيات تعلم الآلة، مثل آلة الدَّعم الموَجَّهة (SVM).
- تقنيات لا تعتمد على وجود معلم، مثل:
  - استنباط المعجم.
  - تقنيات التعلم الذاتي باستخدام التمهيد (Bootstrapping).
  - منهج التعلم المختلط (وجود وعدم وجود معلم).
- تقنيات دلالية، وهي تقنيات تعتمد على تمييز الكلمات وحساب الارتباط الدلالي بينها باستخدام أساليب مختلفة مثل:
  - فهرسة الدلالات الكامنة (Indexing Semantic Latent).
  - أساليب المعاملات الأرجح.
  - أساليب المعلومات المتبادلة نقطة بنقطة (Information Mutual wise Point).
- تقنيات مهجنة تجمع بين نوعين أو أكثر من التقنيات السابقة.

٢,٣- التعرف على طبيعة الكلمات والعبارات اللغوية (كلمات موضوعية بالمقارنة إلى الكلمات التقديرية)

من وجهة نظر التنقيب في الآراء يتم تقسيم نوعية الكلمة إلى نوعين رئيسَين:

- الكلمات الموضوعية (words Objective).
- الكلمات التقديرية (words Subjective) التي تعبر عن الخصائص؛ وتحديد قيمتها في الغالب تقديري.

  قوي - الأفضل

أفقي - سائل

أصفر - أبيض - أسود

تركز البحوث الحالية على التعامل مع الكلمات والعبارات التي يطلق عليها كلمات المحتوى مثل: (الأسماء، الأفعال، الصفات، الحال) وتعتمد هذه الأبحاث على استخدام برمجيات تمييز أجزاء الكلام Part-of-Speech (POS) Tagging ، ويقصد بها تحديد وتصنيف نوع الكلمة: فعل، فاعل، اسم مفرد، اسم جمع، صفة، حال، أداة تعريف، وهكذا. بالنسبة لبعض أنظمة الحاسب مثل فإنها تعرف ما بين ٥٠ إلى ١٥٠ علامة تمييز للغة الإنجليزية وكذلك الحال بالنسبة للغة العربية.

كما تستخدم أساليب لغوية أخري مثل استخراج الجذر وتحديد الجذع للكلمة ومثل حذف الكلمات الوظيفية وغيرها من الكلمات التي تعرف باسم (Stop words) (مثل كلمة بينما).

٢,٤ - تحليل المعنى التقديري للكلمة (Subjectivity Analysis)

يختص تحليل المعنى التقديري للكلمة بتحديد إلى أي الفئات تنتمي قطبية الكلمات: هل هي إيجابية أم سلبية أم حيادية

كلمات إيجابية، مثل: ممتاز - رائع - جيد - بمهارة - متقن

كلمات سلبية، مثل: سيء - حزن - مع الأسف - يتألم

كلمات حيادية، مثل: جدا - كثيرا - قليل - طويل

وقد يبدو للوهلة الأولى أن هذه مهمة سهلة، بالفعل هذا سهل بالنسبة للشخص الذي يتحدث ويتقن اللغة ويمتلك المعرفة البديهية (Commonsense Knowledge) ولكن بالنسبة للحاسب الآلي فهذا يمثل تحديًا كبيرًا له نظرا لتعدد ظواهر اللبس في اللغة ونظرا لعدم إمكانية تغذيته بجميع المعارف البديهية وصعوبة تمييز المعاني الضمنية وقراءة ما بين السطور. فعلى سبيل المثال.. عندما يأتي في النص عبارة «طيب القلب» هل هي عبارة إيجابية أم عبارة سلبية! (بالطبع تخضع لسياق الجملة).

وكما في تقنيات المهام السابقة مثل تقنيات التعرف على الكائِنات الاسميَّة وتقنيات

التعرف على القائم بإبداء الرأي فإن تقنيات مهمة تحليل المعنى التقديري للكلمة تنقسم إلى قسمين رئيسيين وهما:

- تقنيات مبنية على النماذج الإحصائية للغة. وتتطلب عادة إعداد كمية كبيرة من النصوص التي يتم إضافة الحواشي لها وتمييز المعنى التقديري للكلمة بينها بطريقة يدوية. ويبقى دور برمجيات تعلم الآلة لاستخلاص وصياغة نماذج التعرف على المعنى التقديري للكلمة. وهي أيضا مكلفة الإعداد ولكن لا تحتاج خبرة اللغويين بمثل احتياج التقنيات اللغوية.
- تقنيات مبنية على قواعد بيانات معجمية. وهي عادة عبارة عن قواعد بيانات لكلمات اللغة مزودة ببرمجيات تحليل صرفية للتعامل مع الاشتقاقات الصرفية المختلفة للكلمة الواحدة. وهذه التقنيات تعطي جودة عالية في التعرف على المعنى التقديري للكلمة، ولكنها عالية التكلفة في الإعداد وتحتاج إلى أشهر من العمل من قبل اللغويين ذوي الخبرة الحسابية.

وتتراوح دقة التقنيات الحالية في تحديد المعنى التقديري للكلمة نسبة تتراوح بين ٧٨ - ٨٧٪ مما يدل على الحاجة إلى مجهودات بحثية مستمرة للوصول إلى دقة أعلى تتناسب واحتياجات التطبيقات العملية.

٢ , ٥- استخراج الرأي وتصنيف النص

معظم تقنيات استخراج الرأي المتاحة حاليا تأخذ في الاعتبار مهام التعرف على طبيعة الكلمات وتحديد المعنى التقديري لها ولكن توجد اتجاهات بحثية أخرى تصل إلى تحديد الرأي بدون المرور بعملية تحديد طبيعة الكلمات ومعناها التقديري.

على سبيل المثال، توجد أنظمة لتحليل اتجاهات مقالات الرأي عن الأفلام السينمائية Pang et al. (2002) باستخدام التقنيات الإحصائية وتقنيات تعلم الآلة المختلفة مثل تقنيات بايز المبَسَّطة (Bayesian Naïve)، وتقنيات آلات الدَّعم المَوَجَّهة (Support vector machines)، وتقنيات الفوضى القُصوى (Maximum Entropy).

وتعتمد هذه التقنيات على استخراج مجموعة من الخصائص من المقالات ودراسة مئات، بل آلاف من المقالات الَّتي تُعرَفُ طبيعة الرأي لها ومحاولة التعرف على العلاقة

بين طبيعة وقيم الخصائص المختلفة التي تم تحديدها (أو يتم أيضا استنتاجها بواسطة البرمجيات). ويتم تطبيق هذه العلاقات على المقالات الجديدة لتحديد طبيعتها واستخلاص الرأي منها.

نماذج من خصائص المقالات المستخدمة في Pang et al. (2002):

- قوائم الكلمات في النص وأماكن تواجدها في النص.
- علامات تمييز أقسام الكلام.
- أعداد تواتر ثنائيات الكلمات بالنص.
- قوائم الصفات الموجودة بالنص.
- أعلى ٢٦٣٣ كلمة أحادية من حيث تكرار تواجدها بالنص.

وقد بلغت دقة حساب رأي المقالات نسبة بين ٧,٧٨٪ و ٩,٨٢٪.

في حالة التقنيات التي تعمل بالاعتماد على طبيعة الكلمات وقيمها التقديرية كما في شغل Hu and Liu (2004) يتم تحديد رأي الجملة وذلك بحساب نسبة مجموع الكلمات الإيجابية بقيمها التقديرية مقارنة بمجموع الكلمات السلبية بقيمها التقديرية (وذلك لكل كلمات الرأي الموجودة بالجملة)، وقد بلغت دقة حساب رأي الجملة نسبة ٢,٨٤٪.

٣- التنقيب في الآراء واللغة العربية

(استخراج الآراء من المعلومات العربية وشرح السببية في مجال الأخبار)

ملحوظة: هذا الجزء من الفصل الحالي مأخوذ من أعمال قمنا بها بمركز التنقيب في البيانات بجامعة القاهرة.

ما هي المسألة: بناء نظام مُمَكن يستقبل النصوص الإخبارية ويقوم باستخراج الآراء بشأن كيان معين يحدده المستخدم.

على سبيل المثال يوجد النص الإخباري التالي:

عقب تقدم مؤسسيه بطلب إشهاره قانونيا، عقد حزب «مصر الأم» مؤتمره الصحافي الأول في القاهرة، لكن الرفض الجماهيري حال من دون إكمال المؤتمر، ربما لهذا الأمر ما يبرره، فقد جاءت أفكار وبرنامج الحزب غريبة على المتلقي المصري، وجاءت دعاوى العودة إلى القومية المصرية القديمة، واعتبار المصريين غير عرب، وأن اللغة التي تتحدث بها مصر هي لغة مصرية وليست عربية كما زعم محسن لطفي السيد وكيل الحزب.

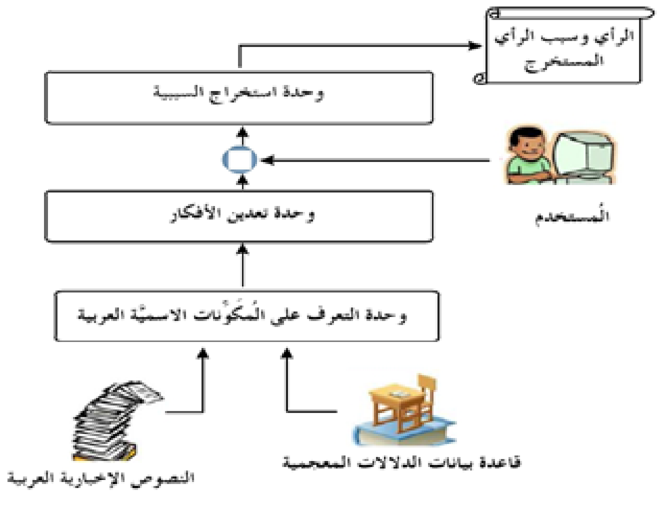
والمطلوب الإجابة على سؤال مثل:

ما رأي الجماهير في حزب مصر الأم؟

على النظام الآلي للتنقيب في الآراء أن يستنتج أن الجماهير لا تحب «حزب مصر الأم» ويستنتج ذلك من تواجد كلمات تحمل مفهوم السلبية في معناها مثل «الرفض الجماهيري» وأن يتعرف على سبب هذا الرأي السلبي. والسبب في حالتنا هذه هو:

«فقد جاءت أفكار وبرنامج الحزب غريبة على المتلقي المصري، وجاءت دعاوى العودة إلى القومية المصرية القديمة، واعتبار المصريين غير عرب، وأن اللغة التي تتحدث بها مصر هي لغة مصرية وليست عربية».

يبين الشكل التالي (الشكل ٤-١١) هيكل نظام استخراج الآراء من المعلومات العربية وشرح السببية له.



الشَّكل ٤-١١: هيكل نظام استخراج الآراء من المعلومات العربية وشرح السببية لها

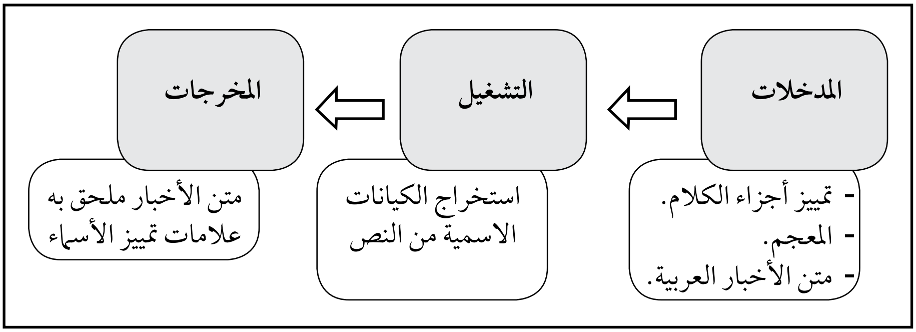
وفيما يلي شرح مبسط لكل وحدة من مكونات نظام استخراج الآراء من المعلومات العربية وشرح السببية له.

٣,١- وحدة التعرف على الكائِنات الاسميَّة العربية

دور هذه الوحدة هو التعرف في نصوص الأخبار على مختلف فئات الكائِنات الاسميَّة العربيَّة (شخص أو منظمة، والموقع والتاريخ والوقت، وأنواع الوظائف، والسيارات والأجهزة والهواتف النقالة، والعملة).

- هذه الخطوة مهمة جدا، لأن هذه الكائِنات الاسميَّة تُمَثِّلُ غالبا أصحاب الرأي الأكثر شيوعا أو كائنات تتصل بها الآراء والأخبار.
- وعلاوة على ذلك، فإن هذه الكائِنات في حد ذاتها هي عبارات موضوعية وليست عبارات دلالية تفيد الرأي وبالتالي وفي وقت لاحق فإن نظام التنقيب والبحث عن الرأي يمكن تجاهل هذه الكائِنات من حيث سلبية أو إيجابية المعنى.

ويُوَضِّح الشكل التالي (الشكل ٤-١٢) كيف تعمل هذه الوحدة:

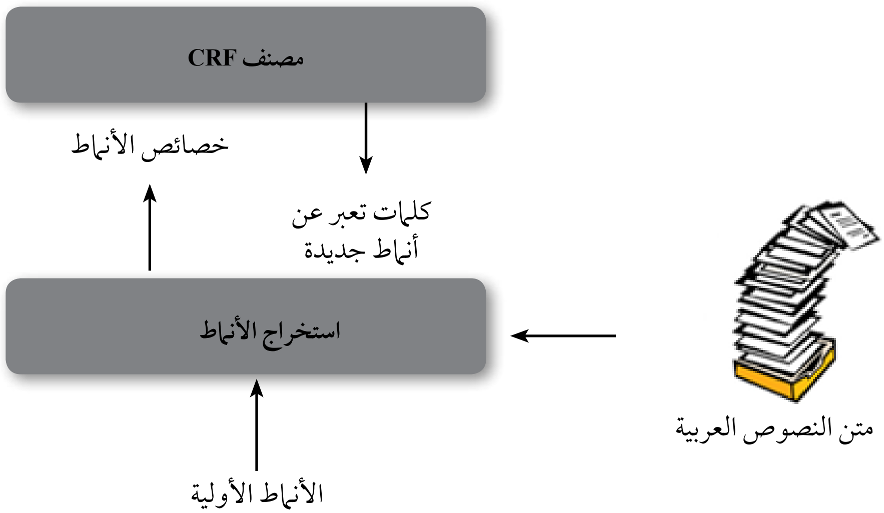


الشَّكل ٤-١٢: وحدة التعرف على الكائِنات الاسميَّة العربية

تستخدم هذه الوحدة مصنف من نوع «الحقول الشرطية عشوائية» Conditional Random Fields (CRF) classifiers الذي سبق الإشارة إليه وذلك للتعرف على الكائِنات الاسميَّة المختلفة.

يعمل هذا المصنف بأسلُوب يعرف باسم (Bootstrapping) وهو أسلُوب للتعلم الذاتي من خلال تزويد المصنف بقائمة أولية من أنماط المسميات المختلفة، مثل: أنماط

الأشخاص أو المنظمات، والمواقع والتواريخ ، وأنواع الوظائف، والسيارات والأجهزة والهواتف النقالة، والعملات النقدية وغيرها. وعلى المصنف زيادة هذه الأنماط كلما تعرض إلى نصوص إخبارية جديدة. ويُوَضِّح الشكل التالي (الشكل ٤-١٣) أسلُوب عمل المصنف من نوع «الحقول الشرطية عشوائية»:



الشكل ٤-١٣: مصنف من نوع «الحقول الشرطية عشوائية».

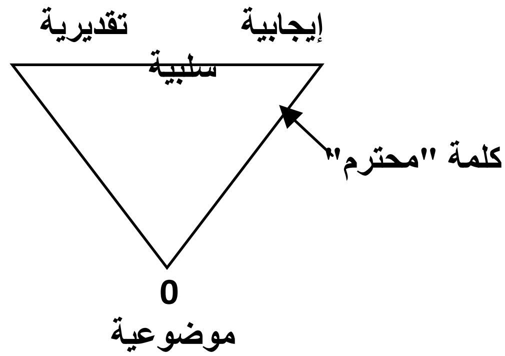
وجدير بالذكر أن هناك مشروعات وأبحاث كثيرة في مجال التعرف على الكائِنات الاسميَّة من النصوص، وقد اتَّجَهَت هذه الأبحاث نحو الدمج بين العلوم الإحصائية وعلوم تعلم الآلة وعلوم اللغويات حتى يمكن أن نصل إلى جودة عالية في التعرف على الكائِنات الاسميَّة. وتمثل عملية التعرف الآلي على الكائِنات الاسميَّة تحديًا كبير نظرا لديناميكية الأسماء وظهور مسميات جديدة ونظرا لقضايا اللبس الناتج من تعدد المعنى المحمول على الكلمات.

٣,٢- وحدة التنقيب عن الرأي

- تشتمل هذه الوحدة على قاعدة بيانات معجمية دلالية لجميع كلمات اللغة العربية حيث توصف كل كلمة بموضوعيتها أو تقديراتها، بمعنى آخر هل هي كلمة موضوعية تقبل الصواب والخطأ مثل كلمة «اليوم» في جملة «اليوم عطلة

رسمية» أم هي كلمة ذات طبيعة تقديرية، مثل كلمة «رائع» في جملة «الطقس اليوم رائع». ومع كل كلمة ذات طبيعة تقديرية توضح قاعدة البيانات المعجمية الدلالية قطبية الكلمة؛ هل تدل على شيء إيجابي أم تدل على شيء سلبي؟. ومن البديهيّ أن توجد كلمات تعبر عن السلبية والإيجابية، وذلك طبقا لسياق الاستخدام. فكلمة «صامت» قد تعني عدم المشاركة والسلبية في الرأي أو قد تحمل المعنى الإيجابي وتعني القدرة على تحمل الموقف (مثال: ظل الشعب صامتا رغم تزوير الانتخابات، ظل الرجل صامتا رغم شدة المرض).

وتوجد جهود كثيفة من فرق أبحاث الشركات العربية العاملة في المجال في بناء قواعد بيانات معجمية دلالية للغة العربية على نمط قواعد البيانات المعجمية الدلالية للغات الإنجليزية (SentiWordNet) حيث تعبر عن قطبية الكلمة (المعنى الشعوري أو العاطفي أو الرأي) بموقعها داخل المثلث كما في الشكل (٤-١٤):



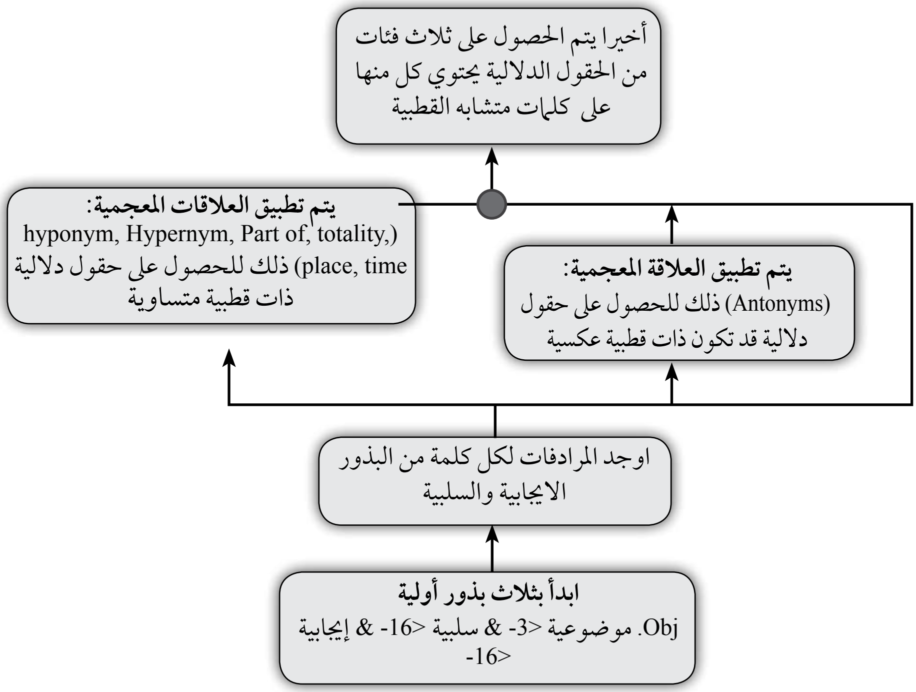
الشَّكل ٤-١٤: قطبية الكلمة

- تقوم الوحدة بتحديد نوعية الكلمة (موضوعية أم تقديرية) وفي حالة التقديرية تقوم الوحدة بتصنيف الكلمة طبقا لقطبيتها.
- استخراج التعبير الرأي من النص وتحديد صاحبه وموضوعه.

أمثلة من قاعدة بيانات معجمية دلالية:

معاني موضوعية	معاني إيجابية	معاني سلبية
• الرجل	• القوة	• سلبي
• الصوت	• الشجاعة	• ردئ
• الآلة	• العطاء	• فشل

ويوضح الشكل التالي (الشكل ٤–١٥) كيفية ترميز كل حقل دلالي بالقطبية المناسبة بأسلُوب (Bootstrapping).



الشَّكل ٤–١٥: التعلم الذاتي باستخدام التمهيد (Bootstrapping)

- بعد ترميز كل حقل دلالي يتم حساب قيم الإيجابية والسلبية والموضوعية لكل كلمة فيه.
- يراعى أن يكون مجموع القيم الإيجابية والسلبية والموضوعية لكل كلمة يساوي رقم الواحد الصحيح.

٣,٣- تصنيف موضوعية النص

لتصنيف الجملة وفقا لتوجهاتها الدلالي، يتم تنفيذ المهام التالية:

- **تحديد القطبية لمكونات الجملة**

يتم تصنيف العبارات في كل جملة. ويطلق على هذه العبارات مصطلح القرائن. والقرينة الواحدة قد تحتوي على أكثر من كلمة واحدة. يتم هذا التصنيف من خلال تحديد العناصر التالية:

- تحديد ما إذا كانت القرينة موضوعية، تقديرية، أم حيادية.
- تحديد اتجاه قطبية القرينة: يميل نحو إيجابية، أم يميل سلبية.
- تحديد قوة قطبية القرينة: درجة الإيجابية، درجة السلبية.

ولكن كيف يتم تحديد القطبية وقوتها لكل قرينة؟

- بالنسبة للقرائن وحيدة الكلمة يتم الحصول على قطبيتها وقوتها من خلال قاعدة بيانات يطلق عليها اسم «المعيار الذهبي» حيث يتم إعدادها يدويا أو يتم إنشائها بأسلُوب (Bootstrapping) المذكور آنفا.
- بالنسبة للقرائن ثنائية الكلمة أو ثلاثية الكلمة أو عدد ن من الكلمات (يطلق على هذا المصطلح «النَّحو العدديّ N-Gram») يتم الحصول على قطبيتها وقوتها باستخدام خوارزم «المعلومات المتبادلة من وجهة النظر النقطية (Pointwise Mutual Information (PMI) Algorithm) وذلك بالرجوع إلى قطبية القرائن وحيدة الكلمة وإلى المعيار الذهبي.

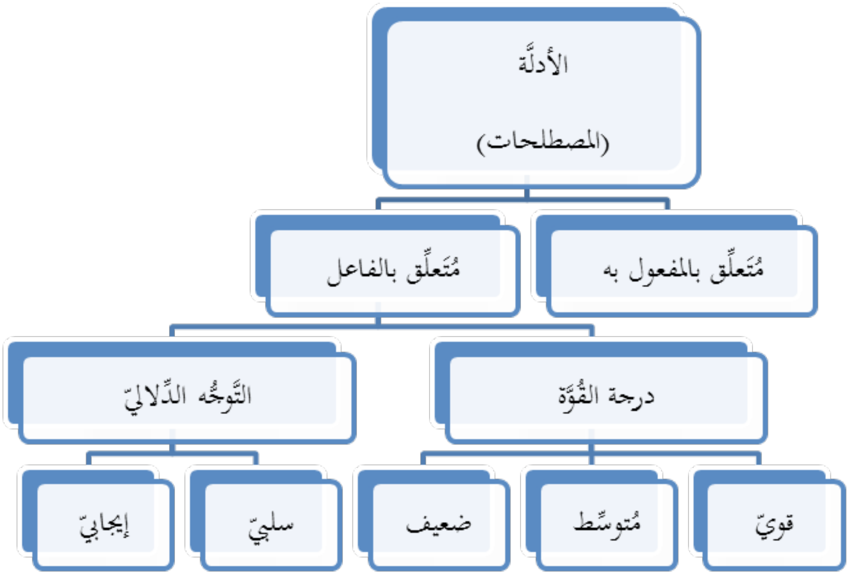
وباختصار شديد فإن قيمة (PMI) بين كلمتين $w1$، $w2$ تعبر عن مدى تواجد هاتين الكلمتين متتابعتين في اللغة العربية:

$$\text{PMI}\ (w1,w2) = \log2\ [\text{p}(w1\&w1)/\text{p}(w1).\text{p}(w)]$$

حيث p تمثل مدى تكرار الكلمة في اللغة، log2 تمثل الدالة اللوغارتمية.

- **تحديد الرأي على مستوى الجملة**

يتم ذلك باستخدام عدة أساليب مختلفة منها ما يعتمد على تطبيق خوارزمات تعلم الآلة ومنها ما يعتمد على استخدام التحليل اللغوي للجملة.



الشَّكل ٤-١٦: تحديد الرأي على مستوى الجملة.

- تحديد الرأي (خوارزمات تعلم الآلة):

تعمل خوارزمات تعلم الآلة على مجموعة من السِّمات يتم استخلاصها من الجملة كما ذُكِرَ في الفصل الثَّاني من هذا الباب (تلخيص النصوص). ويتم التعبير عن هذه الخصائص بدلالة قطبيات القرائن التي تم استخراجها من الجملة.

وتتطلب خوارزمات تعلم الآلة وجود مُدَوَّنة نَصِّيَّة مُعَنونة يدويا بالرأي أو بطريقة (Bootstrapping) لتوفير المجهود اليدوي.

عادة ما تكون نتائج استخدام خوارزمات تعلم الآلة جيدة إذا كانت الجمل الجديدة المراد استخلاص الرأي منها تأتي من نفس مجال المدَوَّنة اللُّغَوِيَّة الَّتي استخدمت في تعليم الآلة. فمن غير المتوقع أن تكون النتائج غير مرضية عند استخدام آلة قد تم تعليمها لنصوص إخبارية في تحليل الرأي لنصوص في مجال الطب.

- تحديد الرأي والتحليل النحوي للجملة العربية:

التحليل النحوي للجملة يتميز بقدرته على الاستخدام في مجالات متنوعة وبالتالي يحل مشكلة خوارزمات تعلم الآلة المذكورة أعلاه.

فالتحليل النحوي الصحيح للجملة يحدد الحدث ومن فاعله ومن وقع عليه ومن قام بالاشتراك فيه، وينتج عنه ربط الصفة بالموصوف، وينتج عنه ربط الضمائر بالكائِنات الاسميَّة الموجودة بالجملة. ولذا يستخدم التحليل النحوي للجملة العربية مقرونا بوحدة التعرف على الكائِنات الاسميَّة لتحديد موضوع الرأي ومن هو صاحب الرأي وما هو الرأي نفسه.

على الرغم من النتائج الإيجابية لوحدة التعرف على الكائِنات الاسميَّة إلا أن خوارزمات التحليل النحوي المتاحة للنصوص العربية لا تضاهي مثيلاتها للغة الإنجليزية وتمثل نقطة الضعف في الوصول إلى أنظمة استخلاص الرأي التي تعمل بكفاءة.

مع زيادة كفاءة المحلل النحوي (والدلالي) ودمجها مع خوارزمات تعلم الآلة فإنه من المتوقع أن ترتفع دقة أنظمة استخلاص الرأي بصورة ملحوظة.

٤- الموارِد اللغوية اللازمة المتاحة والمطلوبة

تحتاج نظم تنقيب الآراء طبقا للتقنيات المستخدمة في تنفيذها إلى أحد أو بعض الموارِد اللغوية التالية:

١- قواعد البيانات المعجمية.

٢- قواعد البيانات المعجمية الدلالية.

٣- نظم المحللات الصرفية والنحوية.

٤- المدَوَّنات النَّصِّيَّة المعَنونة الدالة على موضوع الكلمة وقيمها التقديرية.

ويلاحظ أن نظم المحللات الصرفية والنحوية وقواعد البيانات المعجمية تعتبر قاسم مشترك لتطبيقات لغوية كثيرة، وبصفة عامة فإن كثير من الموارِد اللغوية متاحة

(بمقابل مادي بسيط للباحثين) للغات الإنجليزية واللغات الأوربية الرئيسية واللغات الأسيوية مثل الصينية ولكنه شحيحة على مستوى اللغة العربية.

فعلى مستوى اللغة الإنجليزية نجد:

٤ , ١ - شبكة الكلمات (WordNet) (من جامعة برينستون) وتشتمل على مفردات اللغة الإنجليزية وأمثلة لاستخداماتها ومعانيها المختلفة ومدى شيوع استخدامها والكلمات المرتبطة ببعضها؛ وغيرها من المعلومات. وبيانها الإحصائي كالتالي:

جزء الكلام	بدون تكرار	ترادف الكلمات	المجموع
	Strings		أزواج الكلمات - المعاني
اسم	١١٧٧٩٨	٨٢١١٥	١٤٦٣١٢
فعل	١١٥٢٩	١٣٧٦٧	٢٥٠٤٧
صفة	٢١٤٧٩	١٨١٥٦	٣٠٠٠٢
حال	٤٤٨١	٣٦٢١	٥٥٨٠
المجموع	١٥٥٢٨٧	١١٧٦٥٩	٢٠٦٩٤١

الجدول ٤-١: أعداد الكلمات وفئات الكلمات والمعاني

قسم الكلام	وحيدة المعنى	متعددة المعنى	متعددة المعنى
	الكلمات والمعاني	الكلمات	المعاني
اسم	١٠١٨٦٣	١٥٩٣٥	٤٤٤٤٩
فعل	٦٢٧٧	٥٢٥٢	١٨٧٧٠
صفة	١٦٥٠٣	٤٩٧٦	١٤٣٩٩
حال	٣٧٤٨	٧٣٣	١٨٣٢
المجموع	١٢٨٣٩١	٢٦٨٩٦	٧٩٤٥٠

الجدول ٤-٢: بيانات تعدد المعاني

قِسم الكلام	متوسط تعدد المعنى شاملة الكلمات ذات المعنى الواحد	متوسط تعدد المعنى باستثناء الكلمات ذات المعنى الواحد
اسم	١,٢٤	٢,٧٩
فعل	٢,١٧	٣,٥٧
صفة	١,٤٠	٢,٧١
حال	١,٢٥	٢,٥٠

الجدول ٤-٣: متوسطات تعدد المعاني لأقسام الكلام

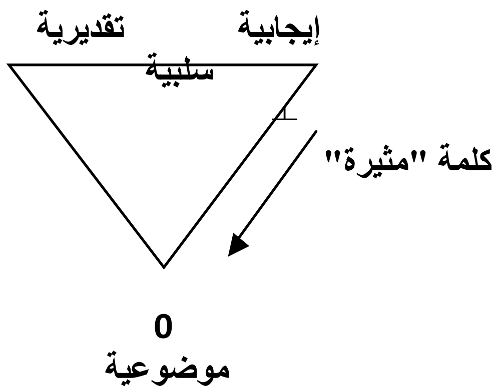
٤,٢ - قواعد البيانات المعجمية الدلالية (SentiWordNet):

وهي شبيهة بشبكة الكلمات (WordNet)، ولكن مع التركيز على معلومات المشاعر (Sentiments) للكلمات. فمع كل مجموعة ترادفات للكلمة يتم وضع البيان الثلاثي التالي (ثلاث أرقام): قيمة الموضوعية، قيمة الإيجابية، قيمة السلبية.

مثال كلمة «interesting» (بمعني مثيرة للاهتمام) تأخذ القيم التالية:

معنى إيجابي: ٠,٣٢٥ معنى سلبي ٠ كلمة موضوعية: ٠

وموقعها داخل المثلث كالتالي:



الشَّكل ٤-١٧: قطبية الكلمة

٤ ,٣- المدَوَّنة اللُّغَوِيَّة المزودة بالحواشي الدالة على موضوع الكلمة وقيمها التقديرية:

ويحتوي على ٥٣٥ مقالة إخبارية تحتوي على ١١١١٤ جملة، كل كلمة فيها مزودة بالحواشي الدالة على موضوع الكلمة وقيمها التقديرية.

٤ ,٤- نظم المحللات الصرفية والنحوية:

يتوافر العديد منها للغة الإنجليزية مثل محلل ستانفورد (من جامعة ستانفورد) ومحلل لينك (من جامعة كارنيجي ميلون).

ومنها من يدعم اللغة العربية مثل محلل ستانفورد ولكن دقة التحليل لم تصل بعد إلى مثيلاتها بالنسبة لتحليل اللغة الإنجليزية.

فمثلا الجملة التالية (تم إدخالها إلى محلل ستانفورد النحوي؛ وهو محلل مبني على النموذج الإحصائي للغة العربية والمدَوَّنات النَّصِّيَّة المعَنونة):

«**مظاهرات أمام مكتب شرف للمطالبة بمحاكمة سميرة مبارك وفريدة حسن.**»

وكان ناتج التحليل اللغوي كالتالي (لاحظ تحليل كلمة «أمام» وكلمة «للمطالبة» وموقعهم في شجرة التحليل وكذلك بالنسبة لحرف «الواو»):

Your query الجملة

مظاهرات أمام مكتب شرف للمطالبة بمحاكمة سميرة مبارك وفريدة حسن.

<u>Tagging الترميز</u>

NNS/ مظاهرات

NN/ أمام

NN/ مكتب

NN/ شرف

DTNN/ للمطالبة

VBP/ بمحاكمة

سميرة/ NNP

مبارك/ NNP

وفريدة/ NNP

حسن/ NNP

./ PUNC

التحليل النحوي Parsing

(ROOT
 (S
  (NP (NNS مظاهرات)
   (NP (NN أمام)
    (NP (NN مكتب)
     (NP (NN شرف)
      (NP (DTNN للمطالبة))))))
  (VP (VBP بمحاكمة)
   (NP
    (NP (NNP سميرة) (NNP مبارك))
    (NP (NNP وفريدة) (NNP حسن))))
  (PUNC .)))

بالطبع مثل هذا التحليل سوف يؤثر على مدى صحة التنقيب.

٥- التوجهات المستقبلية والتحديات الَّتي تواجه تنقيب الآراء

١- بناء قواعد بيانات متخصصة للكائِنات الاسمية (Gazetteers). فمثلاً هناك قواعد تشتمل على أسماء الأشخاص وأخرى على أسماء المدن وهكذا. ويتم ذلك من خلال برمجيات عديدة تحاول تجميع هذه البيانات.

٢- بناء محللات إعرابية عالية الجودة؛ فلا تزال هذه المحللات ينقصها الكثير حتى تستطيع أن تتعامل مع قضايا الالتباس في الجملة وخصوصا في الجمل الطويلة.

٣- تحديد ما إذا كان المستند أو الجزء (الجملة أو الفقرة) ذاتيًّا متعلقا بالفاعل ويعبر عن الرأي.

٤- الصعوبة الَّتي تقع نتيجة ثراء اللغة البشرية.

٥- يمكن أن تعبر كلمة أساسية واحدة عن ثلاثة آراء مختلفة (رأى إيجابيّ، ومتعادل، وسالب بالترتيب) معتمدة على سياق النص.

٦- من أجل الوصول إلى موجز محسوس أو استنتاجات واضحة فإن تحليل وجهات النظر يجب أن يشمل فهم السياق.

٧- معظم الأبحاث في تقنيات التنقيب في الآراء تتعامل مع كلمات المحتوى كما ذكرنا سابقا ولكن هناك أنواع أخرى من الكلمات لها تأثيرها في تحديد الرأي المرتبط بالنص مثل «ومع ذلك» ومثل كلمة « لكن» في النص التالي:

«هذا الكتاب جيد ولكنه صعب الفهم»

طبقا لكلمات المحتوى الموجود بالنص فإن الرأي المستخلص هو «حيادي» نظرًا لأن عدد الكلمات الإيجابية في النص (كلمة جيد) تساوي عدد الكلمات السلبية فيه (كلمة صعب). البعض منا يمكن أن يصنف الرأي في الكتاب بأنه إيجابي نظرا لأن موقع كلمة «جيد» جاءت قريبة من الكتاب. إنما إذا أخذنا في الاعتبار كلمة «لكن» الموجودة في النص فالبعض الأخر من المكن أن يعتبر أن يصنف الرأي في الكتاب بأنه سلبي، وهكذا.

٨- الأخذ في الاعتبار أخطاء الكتابة والقدرة على تصحيح الأخطاء.

٩- التعامل مع ما يمكن تسميته الجمل ذات العلاقات العميقة المتداخلة مثل:

«هذا الطالب توفرت له جميع إمكانيات النجاح من ذكاء وسرعة بديهة وقوة ذاكرة ومهارة مدرس لكن قدر الله نافذ».

«أيها الطالب العبقري! هناك حل أسهل كثيرا».

«ولكن قومي وإن كانوا ذوي عدد ليسوا من الشر في شيء وإن هانا».

١٠- صعوبة التوافق البشري على نفس المستند، فهناك ما يقرب من فرصة ٨٢٪ أن يتفق اثنان أو أكثر من المحللين البشريين مع بعضهم البعض.

١١ .- وغيرها من التحديدات الَّتي تتطلب الكثير من الأبحاث.

ببليوجرافيا مرجعيَّة

المبحث الأوَّل

1. Abu Tair M. M. (2013) Large-Scale Arabic Text Classification: An Approach Towards Distributed Data Mining, Lambert Press.
2. Aggarwal, C. C.; Zhai, C. (2012). Mining Text Data. Springer.
3. Alghamdia, H. M. et al. (2014). Arabic web pages clustering and annotation using semantic class features. Journal of King Saud University - Computer and Information Sciences, Vol 26, Issue 4.
4. Baker, S. (2018). Semantic Text Classification for Cancer Text Mining. University of Cambridge.
5. Billard, L.; Diday, E. (2019). Classification Methodology for Symbolic Data. Wiley.
6. Clarke, B. S.; Fokoué, E.; Hao Helen Zhang, H. H. (2009). Principles and Theory for Data Mining and Machine Learning. Springer.
7. Indurkhya, N.; Damerau, F. (2010). Handbook Of Natural Language Processing, 2nd Edition. Boca Raton, FL: CRC Press.
8. Kaur, R.; Kaur A. (2016). Text Document Clustering and Classification using K-Means Algorithm and Neural Networks, Indian Journal of Science and Technology, Vol. 9, Issue 40.
9. Miner, G.; John Elder, J.; Hill, T.; Delen, D.; Fast, A. (2012). Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications. Academic Press.
10. Sahamim M. (2009): Text Mining: Classification, Clustering, and Applications. CRC Press.
11. Srivastava, A.; Sahami, M. (2010): Text Mining: Classification, Clus-tering, and Applications. CRC Press.
12. Tan, P. (2013): Introduction to Data Mining. Prentice Hall.

13. Berry, M. W.; Kogan, J. (2010). Text Mining: Applications and Theory. John Wiley & Sons.

14. El-Haj, M.; Kruschwitz, U.; Fox, C. (2011). "Experimenting with Automatic Text Summarisation for Arabic", Lecture Notes in Computer Science, volume 6562, pages 490-499. Springer.

15. Foong, O. M.; Oxley, A.; Sulaiman, S. (2010). "Challenges and Trends of Automatic Text Summarization", IJITT, Vol. 1, Issue 1.

16. Gupta, V.; Lehal, G. S. (2009). "A Survey of Text Mining Techniques and Applications", Journal Of Emerging Technologies In Web Intelligence, Vol. 1, No. 1.

17. Hearst, M. (1999). "Untangling Text Data Mining", in the Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics.

18. Hui, E. (2018). Learn R for Applied Statistics: With Data Visualizations, Regressions, and Statistics. Apress.

19. Juan-Manuel Torres-Moreno (2014), Automatic Text Summarization, John Wiley & Sons.

20. Kim, J. (2019). Genome Data Analysis. Springer Singapore.

21. Miner, G.; Elder, J.; Hill. T; Nisbet, R.; Delen, D.; Fast, A. (2012). Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. Elsevier Academic Press.

22. Zizka, J.; Darena, F.; Svoboda, A. (2019). Text Mining with Machine Learning. Taylor & Francis Group.

المبحث الثَّالث

23. AbdelRahman, S.; Elarnaoty, M.; Magdy, M.; Fahmy, A. (2010). “Integrated Machine Learning Techniques for Arabic Named Entity Recognition”. IJCSI International Journal of Computer Science Is-sues, Vol. 7, Issue 4.

24. Binali, H.; Potdar, V.; Wu, C. (2009). “A state of the art opinion mining and its application domains”. 2009 IEEE International Conference on Industrial Technology, Gippsland, Australia.

25. Blokdyk, G. (2018). Text Mining Complete Self-Assessment Guide. Emereo Pty Limited.

26. Bodendorf, F.; Kaiser, C. (2010). “Mining Customer Opinions on the Internet - A Case Study in the Automotive Industry”. 2010 Third International Conference on Knowledge Discovery and Data Mining.

27. Jo, T. (2018). Text Mining: Concepts, Implementation, and Big Data Challenge. Springer.

28. Liu, B. (2010). “Sentiment Analysis and Subjectivity”. Handbook of Natural Language Processing, Second Edition, (editors: N. Indurkhya and F. J. Damerau).

29. Liu, B. (2015). “Sentiment Analysis: Mining Opinion, Sentiment, and Emotions”. Cambridge University Press.

30. Mei, Q. (2011): Contextual Text Mining. Proquest, Umi Dissertation.

31. Pang, B.; Lee, L.; Vaithyanathan, S. (2002). “Thumbs up? Sentiment Classification using Machine Learning Techniques”. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).

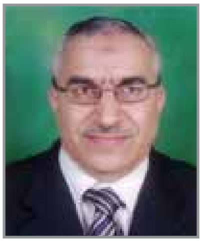
32. Prado, H. A.; Ferneda, E. (2008): Emerging Technologies of Text Mining: Techniques and Applications. IGI Global Snippet.

33. Su, F.; Markert, K. (2008). “From Words to Senses: a Case Study in Subjectivity Recognition”. Proceedings of Coling 2008, Manchester, UK.

34. Weiss, S. M.; Indurkhya, N.; Zhang, T. (2010). Fundamentals of Predictive Text Mining. Springer.

35. Zhao, Y. (2012): R and Data Mining: Examples and Case Studies. Academic Press.

الباحثون



الدكتور/ مُحسِن عبد الرَّازق علي رشوان

يشغل منصب أستاذٍ بقسم الإلكترونيات والاتصالات الكهربائية في كُلِّيَّة الهندسة - جامعة القاهرة. تخرَّجَ عام ١٩٧٧ وكان الأول على دفعته، وحصلَ على ثلاثة ماجستيرات، ثم على الدُّكتوراه من جامعة كوين بكندا؛ أشرف على أكثر من مائة رسالة ماجستير ودكتوراه. يدير الشَّركة الهندسيَّة لتطوير النُّظُم الرَّقمِيَّة RDI المتخَصِّصة في مجال تقنيات اللُّغة العربيَّة.



الدكتور/ المُعتزّ بالله السَّعيد طه

أستاذ الدِّراسات اللُّغويَّة الـمُساعد بجامعة القاهرة، وأستاذ اللِّسانيَّات الحاسوبيَّة المـُشارك بمعهد الدَّوحة للدِّراسات العُليا، ومُنسِّق وَحدة الموارد الـمُعجميَّة بمشروع مُعجم الدَّوحة. نَشَرَ نحوَ ثلاثينَ ورقة علميَّةٍ، بالإضافة إلى عددٍ من الكتب في الـمُعجَمِيَّة العَرَبيَّة والدِّراسات اللُّغَوِيَّة الـمُعاصِرة، وأسهمَ في أكثر من عشرة مشرُوعاتٍ بحثيَّةٍ دوليَّةٍ في ميادين مُعالَجَة اللُّغات الطَّبيعيَّة. حصلَ على عددٍ من الجوائز في مَيدان تخصُّصه، منها: جائزة (ألكسو ALECSO) للإبداع والابتكار في «المـَعلُوماتيَّة والـمُعالَجَة الآليَّة للُّغة العربيَّة»، وجائزة راشد بن حُمَيد للعُلُوم والثَّقافة.



الدكتور/ أسامة إمام

حصل من جامعة القاهرة على بكالوريوس الهندسة الحَيوِيَّة الطِّبِّيَّة والمنظومات عام ١٩٨٤م، وعلى درجة الماجستير عام ١٩٨٧م، ثم على درجة الدكتوراه في ذات التَّخَصُّص عام ١٩٩٠م. يعملُ - في الوقت الحالي - مُديرًا لمركز أبحاث الذكاء الأصطناعي بشركة IBM - مصر. نَشَرَ أكثرَ من ٥٠ ورقة عِلمِيَّة حولَ حوسبة اللُّغة العَرَبيَّة وتقنياتِها في دوريَّاتٍ علميَّةٍ ومُؤتمراتٍ دَوليَّة مُتَخَصِّصة. سَجَّل ٣١ براءة اختراع؛ وحصلَ على العديد من الجوائز العِلمِيَّة.



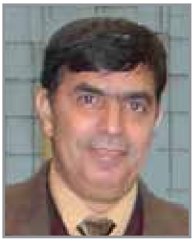
الدكتور/ وَلِيد مجدي

أستاذ مساعد في جامعة إدنبرة ببريطانيا وزميل في معهد ألان تيورينج في لندن. يحمل درجة الدكتوراه من جامعة دبلن في أيرلندا في عُلُوم الحاسب ودرجتي الماجيستير والبكالوريس من كلية الهندسة جامعة القاهرة؛ وهو مُختصّ في مجال الحوسبة الاجتماعية واسترجاع المعلومات. له أكثر من ٦٠ ورقة عِلمِيَّة في دوريَّاتٍ علميَّةٍ ومُؤتمراتٍ دَولِيَّةٍ مُتَخَصِّصة؛ ولهُ تسع براءات اختراع مُسَجَّلَة باسمه في أوروبا والولايات المُتَّحِدة الأمريكيَّة. عَمِلَ في عددٍ من الشَّركات والمؤسَّسات العلميَّة، منها: مايكروسوفت و IBM ومؤسسة قطر.



الدكتور/ أحمد رافِع

حصلَ على درجة الدُّكتوراه من جامعة بول ساباتييه في تولوز بفَرَنسا؛ ويَعملُ أستاذًا لعُلُوم الحاسب بالجامعة الأمرِيكِيَّة في القاهِرة. شارك – باحثًا رئيسًا – في العديد من المشروعات الدَّولِيَّة المَعنِيَّة بتطوير التَّرجمة الآليَّة والتَّنقيب عن الآراء في شبكات التَّواصُل الاجتماعيّ والتَّنقيب في نُصُوص اللُّغة العَرَبِيَّة؛ وتعاوَنَ – في هذه المشرُوعات – مع جامعاتٍ ومؤسَّساتٍ بحثيَّة في أوروبا والولايات المُتَّحِدة الأمريكيَّة.



الدكتور/ علي علي فهمي

هو العَمِيدُ السَّابق لكُلِّيَّة الحاسبات والمعلومات في جامعة القاهرة؛ يعمل – في الوقت الحاليّ – أستاذًا في الذَّكاء الاصطناعيّ وتعلم الآلة. عملَ خلال الفترة من ٢٠٠٥ إلى ٢٠١٠ مُديرًا لمركز التميز في التَّنقيب في البيانات ونمذَجة اللُّغة DMCM في مِصر، وله إسهاماتٌ بحثيَّةٌ بارزةٌ في تقنيات اللُّغة العَرَبِيَّة وتطبيقاتِها.

مباحث لغوية ٥٧

المُعالَجة الآليَّة للنُّصوص العربيَّة

يُصدِر مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية هذا الكتاب ضمن سلسلة (مباحث لغوية)، وذلك وفق خطة عمل مقسمة إلى مراحل، لموضوعات علمية رأى المركز حاجة المكتبة اللغوية العربية إليها، أو إلى بدء النشاط البحثي فيها، واجتهد في استكتاب نخبة من المحررين والمؤلفين للنهوض بعنوانات هذه السلسلة على أكمل وجه.

ويهدف المركز من وراء ذلك إلى تنشيط العمل في المجالات التي تُنَبّه إليها هذه السلسلة، سواء أكان العمل علميا بحثيا، أم عمليا تنفيذيا، ويدعو المركز الباحثين كافة من أنحاء العالم إلى المساهمة في هذه السلسلة.

وتودّ الأمانة العامة أن تشيد بجهد السادة المؤلفين، وجُهد مُحرِّرَي الكتاب، على ما تفضلوا به من رؤى وأفكار لخدمة العربية في هذا السياق البحثي.

والشكر والتقدير الوافر لمعالي وزير التعليم المشرف العام على المركز، الذي يحث على كل ما من شأنه تثبيت الهوية اللغوية العربية، وتمتينها، وفق رؤية استشرافية محققة لتوجيهات قيادتنا الحكيمة.

والدعوة موجّهة إلى جميع المختصين والمهتمين للتواصل مع المركز؛ لبناء المشروعات العلمية، وتكثيف الجهود، والتكامل نحو تمكين لغتنا العربية، وتحقيق وجودها السامي في مجالات الحياة.

الأمين العام للمركز

أ.د. محمود إسماعيل صالح

مركز الملك عبدالله بن عبدالعزيز الدولي
لخدمة اللغة العربية
King Abdullah Bin Abdulaziz Int'l Center for
The Arabic Language

ص.ب ١٢٥٠٠ الرياض ١١٤٧٣
هاتف:٠٠٩٦٦١١٢٥٨٧٢٦٨ - ٠٠٩٦٦١١٢٥٨١٠٨٢
البريد الإليكتروني: nashr@kaica.org.sa

9786038221525